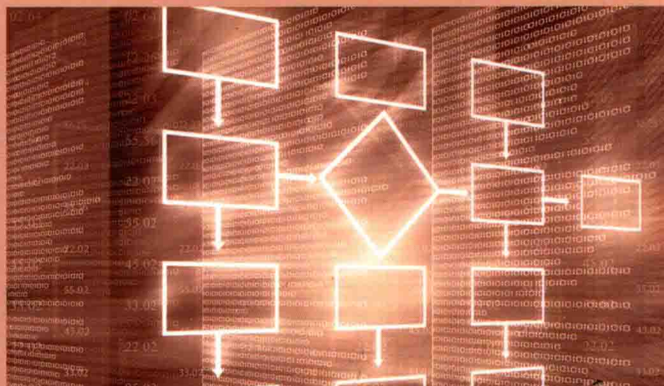
格致方法 · 社会科学研究方法译丛



陈硕 译

Gary King, Robert O. Keohane, Sidney Verba

[美] 加里·金 罗伯特·基欧汉 悉尼·维巴 著

社会科学中的研究设计

Designing Social Inquiry

Scientific Inference
in Qualitative Research

格致出版社 上海人民出版社

世纪出版

也许像我们这样研究比较政治及国际关系的学者最终会感到孤立。这样说的原因是我们无法用统计方法来研究问题并分析数据。很多时候，研究者创造出一些想当然且怪异的应对策略，但并没有系统评估不同方法间的优缺点。金、基欧汉及维巴在本书中给出了实证推论涉及一系列基本问题的定义，并指出，采用定性方法进行研究时所面临的问题其实和定量研究遇到的问题极为相似。作者们也给出了处理这些问题的方法。在本书中，他们为整个社会科学建立了一个普遍适用的架构。我强烈建议立志于社会科学研究的学生掌握这本书的内容。

——罗伯特·贝茨，行为科学高等研究中心

国际关系领域学者的必备读物。

——布鲁斯·拉西特，耶鲁大学

在这本书还是初稿的时候，我就在自己的课堂上使用它。学生们对本书给予了很高的评价。这是一本非常精彩的著作。对于如何结合数据从事研究，作者在书中提供了很多具有可操作性的建议。

——戴维·罗德，佛罗里达大学

上架建议：社科研究方法

ISBN 978-7-5432-2426-1

9 787543 224261 >

定价：38.00元
易文网：www.ewen.cc
格致网：www.hibooks.cn

您可以通过如下方式联系到我们：
邮箱：hibooks@hibooks.cn

微信　微博

格致方法·社会科学研究方法译丛

陈硕 译

Gary King, Robert O. Keohane, Sidney Verba

[美] 加里·金 罗伯特·基欧汉 悉尼·维巴 著

社会科学中的研究设计

Designing Social Inquiry

Scientific Inference
in Qualitative Research

格致出版社 上海人民出版社

图书在版编目(CIP)数据

社会科学中的研究设计/(美)金(King, G.),(美)基欧汉(Keohane, R.O.),(美)维巴(Verba, S.)著;陈硕译.—上海:格致出版社:上海人民出版社,2014

(格致方法.社会科学研究方法译丛)

ISBN 978-7-5432-2426-1

Ⅰ.①社… Ⅱ.①金… ②基… ③维… ④陈… Ⅲ.①社会科学-研究方法 Ⅳ.①C3

中国版本图书馆CIP数据核字(2014)第149581号

责任编辑 顾 悦
装帧设计 路 静

格致方法·社会科学研究方法译丛

社会科学中的研究设计

[美]加里·金 罗伯特·基欧汉 悉尼·维巴 著
陈硕 译

出 版	世纪出版股份有限公司 格致出版社 世纪出版集团 上海人民出版社 (200001 上海福建中路193号 www.ewen.cc) 格致出版 编辑部热线 021-63914988 市场部热线 021-63914081 www.hibooks.cn	印 刷	上海市印刷十厂有限公司
		开 本	787×1092 1/16
		印 张	15.5
		插 页	1
		字 数	219,000
		版 次	2014年9月第1版
发 行	上海世纪出版股份有限公司发行中心	印 次	2014年9月第1次印刷

ISBN 978-7-5432-2426-1/C·109 定价:38.00元

加里·金

哈佛大学政府学院阿尔伯特·魏斯赫德三世学院教授，曾是美国国家科学院、美国统计学学会、美国科学促进会、美国艺术及科学学院、政治学方法论学会以及美国政治及社会科学学院的院士或成员。研究领域为社会科学方法论。

罗伯特·基欧汉

普林斯顿大学伍德罗·威尔逊学院教授，曾出任美国国际关系学会及美国政治科学学会的会长，并是美国艺术及科学学院以及国家科学院院士，也是美国哲学学会的成员。研究领域为国际关系。

悉尼·维巴

哈佛大学政府学院荣休教授，曾出任该校政府系系主任、政府学院副院长等行政职位，也曾担任哈佛大学出版社董事会主席及美国政治科学学会会长。他也曾是美国国家科学院以及美国艺术及科学学院院士、美国哲学学会成员。研究领域为美国研究及比较政府研究。

陈硕

2011年毕业于香港科技大学社会科学部，获社会科学博士学位。目前是复旦大学经济学院教师。主要研究领域为中国的发展经济学、政治经济学及经济史。

译 者 序

还是学生的时候，在论文研讨会上听到坐在下面的教授经常说“你的结果高估了真实作用大小”、“你的解释变量存在内生性问题”、“你的因果关系可能只是相关”、“样本也有自选择问题”。这时候，瞧陈述者那如坐针毡、面红耳赤、结结巴巴的样子，坐在旁边的我顿时心花怒放：“这太过瘾了，我也要做这样的教授，如此评点别人的研究！”但求知欲和进取心只是第一步，除此之外还需要一个好的老师领你入门。我想说的是，《社会科学中的研究设计》(以三位作者姓氏首字母及出版时间构成的简称“KKV1994”也许在学界更具知名度)正充当这个角色。这本书是我在香港科技大学社会科学部硕士阶段的必修课《社会科学方法论》的指定教材。通过一个学期的学习，我发现原来教授们的秘密都藏在这本书中。

《社会科学中的研究设计》由普林斯顿大学出版社于1994年出版。该书出版后就成为众多欧美高校经济学、政治科学及社会学等学科本科及研究生方法课程的指定参考书，也是当今最具影响力的社会科学方法论著作。这也是我读的第一本方法论著作，对此我感到非常幸运。在2006年的时候我便萌发了翻译本书的念头，期间经历了硕士毕业、博士毕业及踏上工作岗位，断断续续竟然用了八年时间完成。大体而言，这本书的翻译经历了四个阶段。第一阶段是“翻译体”阶段：按照原文翻译成中文，不考虑中文表述习惯；第二阶段是润色阶段：结合原文及译文重新表述，使之符合中文习惯；第三阶段是修正阶段：通读译文找到翻译错误及逻辑表述不清的地方，再结合原文修正；第四阶段是中文润色阶段：完全离开原文，按照中文表述重新润色译文。我一直认为

译文符合中文表述习惯非常重要,这可以更有效地帮助读者掌握书中的知识及作者的观点。

就本书内容来说,作者在第1章便明确指出该书的写作目标是“向读者系统地介绍如何在社会科学领域开展研究”。而对于研究实施的两个方法:定量传统及定性传统,作者认为它们之间的差别仅仅是风格和方法论上的不同,所依赖的逻辑都是相同的,“即怎样提出研究问题、开展学术研究并进行有效的描述及因果推论”。不论采用何种方法,评价其研究过程是不是科学的主要标准都要看其是否满足:(1)以推论为研究目的;(2)采用公开的研究程序;(3)结论是不确定的;(4)研究的内容是关于方法的。围绕上述这些观点,本书剩余部分进一步由两个方向展开:第一个方向将研究过程分为描述性推论(第2章)及因果性推论(第3章)两个步骤,其中前者是研究的起点并构成后者的基础。在作者看来,这两个步骤均存在推论的过程:描述性推论涉及如何由样本推及整体,而因果推论则试图发现不同因素间真实的因果关系。本书第2章及第3章提供了这两个推论的定义及推论规则。第二个方向为主题研究,作者向读者提供处理研究中经常遇到的特定问题的对策。这些问题包括:如何确定观测对象(第4章)、如何增加观察值数量(第6章)及如何避免某些特定问题(测量误差、遗漏变量、无关变量及内生性等,见第5章)。就如何确定观察对象来说,作者指出在大多数定性研究中随机选择样本方法具有一定的局限性,在这种情况下有目的地选择样本也许更为可取。围绕有目的取样,作者提供了五种可供读者参考的操作方案。就研究中需要避免的问题来说,测量误差、遗漏变量、无关变量及内生性问题要么导致推论效率降低,要么导致更严重的有偏推论。这些问题在定量研究中尤为突出,值得读者重视。辅之以具体的例子,作者阐述这些问题造成的上述后果并提供了对策。而最后一章讨论的增加样本数量则是定量及定性学者面临的共同问题,如何在现有限制下尽量拓展样本数以提高推论效率的方法会在这一章给出。

严谨的因果思维能力是高度训练后养成的自觉。对于初学者来说,这本书提供了训练所必需的知识,同时又维持在一个恰当的难度水

平之上。作者们主要借助浅显易懂的文字表述而非抽象复杂的模型推导告诉读者在研究设计中会遇到的问题、造成的后果及相应的对策。在维持读者信心的同时,循序渐进地揭示出描述性推论及因果推论的相关知识。本书的风格使我从没感到方法论很枯燥。恰恰相反,它是我认为最有意思的一门学问。同时,对于学者来说,这也是一本常读常新的著作。工作以后,我依然经常抽空翻翻它。书中的内容不断提醒着我在研究过程中要避免作者们提到的那些问题。这种互动,正如作者在中文版前言中所写:"本书的信息可以促进读者重新审视自己的研究,同时在有了实际研究经验之后重读本书还能获得新的感悟。"

作为译者,我很荣幸在这本书出版20周年的时候把它介绍给中国读者。希望它的内容可以对大家的研究工作起到实实在在的帮助。当然这本书的读者绝不局限于学生和学者,政策制定者及业界分析人士一样会从中受益。本书的翻译得到了2014年度"教育部哲学社会科学研究后期资助项目(项目号:14JHQ018)"的资助。最后,我要感谢格致出版社引进本书的中文版权,也感谢为本书顺利出版付出时间和精力的李娜、王亚丽及顾悦三位编辑。对于翻译过程中存在的任何错误及不足之处,希望读者朋友与我联系,我们会在新的版本中予以修订。

陈　硕

复旦大学经济学院

中文版序

政治学用科学方法研究政治现象比社会学起步晚，直到1986年，旨在推动量化研究的政治学方法分会(Society of Political Methodology)才成立。但其后，政治学在对方法的运用上突飞猛进，良作辈出，其运用科学方法的深度和广度已超过了社会学。如今，比较政治学、定性与多元方法研究、政治学方法已成为美国政治学学会(American Political Science Association)中成员规模最大的三个分会。政治学对研究方法的重视程度由此可见一斑。

政治学研究方法的科学化与加里·金教授的贡献分不开。加里·金是哈佛大学政府学院的教授，该校授予了他荣誉最高的大学教授(university professor)头衔。他担任过美国政治学学会方法分会的会长(1997—1999年)，还担任过美国政治学学会的副会长(2003—2004年)，2010年当选美国科学院院士。金迄今为止发表过130多篇论文和8本专著，其中绝大多数成果都是对社会科学研究方法的发展和应用。他对研究方法的创新不仅在理论上，还在应用上。他提出的从汇总数据推断个体行为的方法、在调查研究中实现跨文化可比性的方法已被广泛采纳于研究和实践领域。他在墨西哥主持了目前规模最大的针对健康政策的随机干预试验——这是政策评估和社会科学领域因果推断中最前沿的研究设计。金的成就不仅仅是他个人的，他对方法论的思考和对统计方法的创新推动了整个学科的发展。他与基欧汉和维巴合著的这本《社会科学中的研究设计》就集成了他在方法论和研究设计上的灼见。

虽然这本书是重点介绍运用在政治学领域的研究设计，但这本书

的应用性不局限于政治科学。虽然这本书是以定性研究设计为主题，但这本书讨论的问题对定量取向的学者同样适用。事实上，许多社会科学分支在研究方法上是共通的，定性方法和定量方法在推论逻辑上也是相通的——这是金所持的基本态度和重要观点。过去，甚至包括现在，无论是在政治学还是在社会学，无论是在北美还是在中国，定性和定量取向的学者常常站在不同的阵营，对彼此的工作互不了解，对孰优孰劣争执不休，在各自研究风格的发展上渐行渐远。但这本书就明确整合了这两种方法的共性，即两者都是“系统和科学的工作”，目的都是要提供“有效的描述性或因果性推论”。而且，此书进一步提出，如果定性研究者要作出更为普遍、逻辑关系更清晰、更确证可靠的推论，如果又能突破数据收集的限制，就应该使用定量研究的科学推论原则和建模方法。这一观点虽然未必能被所有学者接受，但却能把研究者的注意力从对风格的争论转移到规范的研究设计上。

规范性是这本书试图传达的另一个观点。相比游记、传记、新闻评论、纪实文学这些同样记述社会事实并表达观点的作品，社会科学研究更强调要采用规范的、训练有素的思维方式，这也是为什么要成为社会科学家通常须接受许多年专业的研究训练，而不是单凭才华或热情就能实现。这本书对规范性的强调体现在它对研究设计提出的具体要求上。例如，书中强调研究问题的选择不仅要有现实价值，更要符合社会科学知识的框架，要能提炼出能够系统性描述或因果推论的具体研究计划。又如，书中反复强调要承认和汇报推论和理论的不确定性，规范的科学研究不应该回避不确定性，虽然难以获得确定无疑的知识，但应该尽可能去评价每一步推论的确定性/不确定性的程度，这才是专业的研究者应当遵循的态度和做法。

设计在研究中的重要性也是本书强调的重要观点，这也是我一直秉持的观点。研究关注的问题大与小、获取的数据多与少，这些固然影响研究的优劣，但研究的灵魂在于设计。一个好的设计能帮助研究者洞悉问题的关键，在数据有限的情况下做出有效的推论。这本书将研究设计分为四个部分：研究问题、理论、数据和对数据的使用。它的第2、3章是对前两个部分作出指导，它的第4、5、6章是对后两个部分提

出建议。这本书对如何利用可观察的现象将理论和数据联系起来、如何提高解释效力、对不确定性的处理等方面提供的方法都颇有启发性和可操作性。此外，这本书对因果关系、遗漏变量偏误、内生性、观察对象的选择等研究领域关注的重要问题都做了专门探讨。

我很高兴看到陈硕博士将此书译成中文，方便广大中国学生和学者了解和学习社会科学的研究方法，从而推动中国社会科学研究的规范化、科学化、专业化。我曾在许多场合提到，中国正经历一场对未来影响深远的社会变迁，这场变迁给当代社会科学家提供了千载难逢的研究机遇——去观察、记录和理解。在这个契机面前，我们更要借助科学、严谨的设计和方法让观察洞悉无遗，让记录客观准确，让理解深刻透彻。我向读者推荐此书，望读者从中受益。

谢　宇
密歇根大学社会学系

中文版前言

在某种程度上，社会科学方法参考书更像一本烹饪指南。虽然书中都告诉读者如何操作，但对其判断并不取决于作者的思维及文字描述是否精确，而是取决于读者最终成果的质量。换句话说，参考书的价值依赖于读者如何使用它们，如何按照书中的提示处理遇到的各种问题。我们都知道，好厨师在烹饪的时候不会完全按照指南中的内容操作。虽然做饭和研究有其各自的原则，但对两者而言，创新都是不可或缺的。好的研究方法参考书能够指导研究者在研究设计中采用合适的分析方法、避免易犯的错误，并在不同的方法间做出取舍。

为了实现上述目标，我们撰写了《社会科学中的研究设计》。当然，我们也希望这本书能够给读者的研究带来灵感并提高他们的能力。这本书不光可以作为研究生课程的教材，也可以为选修课程的学生们提供解决实际问题的建议。国际关系学、政治科学及公共政策的博士课程也经常使用本书作为教材。我们发现学生从本书中受益良多：本书的信息可以促进学生重新审视自己的研究，在有了实际研究经验之后重读本书还能获得新的感悟。读者应当知道，理论学习当然必不可少，但研究过程中的学以致用则最为重要。

就课程设置来说，我们一般在头几节课介绍本书的前言部分，同时也会提一下书中的相关参考文献。这些内容将被放置在以托马斯·库恩(Thomas Kuhn)、伊姆雷·拉卡托斯(Imre Lakatos)等学者为代表的科学哲学范畴内加以理解。完成这些之后，学生会结合本书内容提交短论文。在接下来的几周时间，学生将结合本书第 2 章及第 3 章中的描述性推论及因果性推论去思考那些他们认为重要的学术论文。这些论文都

是相关领域最好的研究，而不是那些很容易就能评判优劣的研究。在这之后，学生将被要求用几周时间写一篇关于测量或过程追踪的课堂论文，以保证他们真正理解了描述性推论。我们有时也鼓励学生重新组织他们在其他课程的课堂论文、学期论文甚至是他们的毕业论文。

课程最后几节课一般是最重要的。这几节课将留给学生做研究设计及陈述。学生先把研究大纲发给其他学生，然后在课堂陈述的时候回应其他学生的点评。这几节课才是本书真正试图传达的精神：怀疑、开放的心态及建设性。本书读者在阅读过程中应当会发现我们很少尖锐地批评其他学者的研究。更多情况下，我们会以那些优秀的研究为例，告诉大家如何在现有基础上通过我们给出的建议进一步提高研究质量。那些提交研究设计的学生将有充足的时间来研究其他学生的意见并思考如何回应。相应地，学生提交的学期论文应当包括针对这些意见的回应及修改。

和好的厨艺相比，好的研究并没有固定套路。仅掌握这本书所传达的观点并不能让一个读者成为一个合格的社会科学研究者。读者应当真正了解本书的内容，将他们和自己的研究相结合并在新研究中加以运用。这种结合必然能让他们的研究更上一层楼，对此我们抱有十足的信心。这本书应当被视为一些重要研究问题的补充，能为大家进一步理解这些议题提供帮助，同时也为以新的视角重新审视这些问题提供基础。就像一本好的烹饪书籍，这本书能够为研究所需的想象力及才智提供广阔的空间。

我们感谢陈硕博士将这本书译介给中国读者。我们希望每一位读者都能被本书激发出新的研究思路，无论是关于经济学、政治学还是社会学。我们也希望读者同样认同这是一本优秀的著作，并把它分享给你们的同事及学生。我们热切期望了解大家的发现，同时也希望本书能够真正地发挥作用。

加里·金

罗伯特·基欧汉

悉尼·维巴

英文版前言

在本书中,我们试图为定性研究中的描述及因果推论构建一个通用的方法,由此建立和定量研究共通的基础。虽然定性研究中涉及的因素在绝大多数情况下是无法定量测量的,但我们仍然认为其背后的基本逻辑和定量研究是相似的。而利用本书所提供的方法,两种研究传统中的不同表述方式就具备了共通的基础。

写作本书的目的是促进定性研究者做出更为科学的因果推论。我们希望这个共通的方法可以为社会科学领域的定性研究者提供帮助并提高其研究质量。因此,本书的内容可以为政治科学家、其他社会科学领域的学者以及业界人士提供参考。其潜在受众既包括定性研究学者也包括统计分析从业者,既包括高年级的本科生及研究生也包括专业学者。为了将定性研究方法更加概念化,我们在书中也借助了一些数理公式。这些公式大都浅显易懂,不需要太多的数学及统计学知识,读者即使略过不看也不影响对本书内容的理解。

大多数学者认为教学和研究是相互补充的。虽然它们之间有少许的差别,但更多的是一种相互协同关系:两者都需要探索和传播新知识。在一定程度上,这本书正是这种协同关系的产物。1989 年我们在哈佛大学政府管理系开设"社会科学中的定性方法"这门研究生课程时就开始筹划这本书。这门课的课堂讨论非常热烈,以至于下课后学生们依然在为某些问题争论不休。在学生和老师的备忘笔记中依然可以找到当时辩论的议题。课堂讨论虽然激烈,但又相当友善。我们在课堂上立的规则就是鼓励分歧而拒绝妥协:哪怕还有一个人没被其他人说服,辩论就不能结束。这门课程结束之后,无论是做定性研究的学生

还是做定量研究的学生都从对方那里学到了丰富的知识，也在很大程度上改变了一些先入为主的看法。在试图为两种风格的研究建立一个共通基础外，书中的内容也被视作我们就如何在定性研究中作出科学的因果推论这一问题上所持的基本观点。这本书的初稿在1991年完成，之后也经历了一系列较大的修改。加里·金是这本书的策划人，他撰写了绝大部分章节的初稿，同时在后续的写作过程中依然担当着主要角色。由于本书大多数内容均由我们三人共同撰写，因此要分清楚某个具体章节的执笔人已经很困难了。

这本书完成之后，我们将初稿送给美国高校的很多学者征求意见，他们给了我们很多高质量的反馈意见。作为教材，哈佛大学和其他高校的学生使用了该书的初稿。他们的反馈和互动对后续的修订工作也起到了积极作用。当然，罗列出所有对本书提出过意见的人是很困难的事情（根据我们的估计，那些给过重要意见的人却被遗漏掉的概率是0.92），我们依然想对以下学者致以特别的感谢：Christopher H.Achen, John Aldrich, Hayward Alker, Robert H.Bates, James Battista, Nathaniel Beck, Nancy Burns, Michael Cobb, David Collier, Gary Cox, Michael C.Desch, David Dessler, Jorge Domínguez, George Downs, Mitchell Duneier, Matthew Evangelista, John Ferejohn, Andrew Gelman, Alexander George, Joshua Goldstein, Andrew Green, David Green, Robin Hanna, Michael Hiscox, James E.Jones, Sr., Miles Kahler, Elizabeth King, Alexander Kozhemiakin, Stephen D. Krasner, Herbert Kritzer, James Kuklinski, Nathan Lane, Peter Lange, Tony Lavelle, Judy Layzer, Jack S.Levy, Daniel Little, Sean Lynn-Jones, Lisa L.Martin, Helen Milner, Gerardo L.Munck, Timothy P.Nokken, Joseph S.Nye, Charles Ragin, Swarna Rajagopalan, Shamara Shantu Riley, David Rocke, David Rohde, Frances Rosenbluth, David Schwieder, Collins G. Shackelford, Jr., Kenneth Shepsle, Daniel Walsh, Carolyn Warner, Steve Aviv Yetiv, Mary Zerbinos 和 Michael Zürn。我们也感谢 Steve Voss 为这本书制作索引。感谢普林斯顿大学出版社的 Walter Lippincott, Malcolm DeBe-

voise，Peter Dougherty，Alessandra Bocco 四位编辑。加里·金感谢国家科学资金(SBR-9223637)的研究项目资助。罗伯特·基欧汉感谢约翰·西蒙·古根海姆(John Simon Guggenheim)纪念基金会的经费资助。

本书的初稿也在众多学术研讨会上被讨论过并获得了同行们的宝贵意见。这些会议包括美国中西部政治科学年会(芝加哥，1990 年 4 月 2—6 日)、政治科学方法论组会(杜克大学，1990 年 7 月 18—20 日)、美国政治科学年会(华盛顿特区，1991 年 8 月 29 日—9 月 1 日)、社会科学哲学方法论研讨会(哈佛大学，国际事务中心，1992 年 9 月 25 日)、统计应用跨学科联合研讨会(印第安纳大学，1992 年 12 月 4 日)、全球合作与变迁研讨会(加州大学伯克利分校，1993 年 1 月 15 日)以及伊利诺伊大学香槟分校的研讨会(1993 年 3 月 18 日)。

加里·金
罗伯特·基欧汉
悉尼·维巴
马萨诸塞州，剑桥

Designing Social Inquiry: Scientific Inference in Qualitative Research
Copyright © 1994 by Princeton University Press.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the Publisher.

This authorized Chinese translation edition copyright © 2014 by TRUTH & WISDOM PRESS.
All rights reserved.

上海市版权局著作权合同登记号:图字 09-2012-868

目　录

译者序　/Ⅰ

中文版序　/Ⅳ

中文版前言　/Ⅶ

英文版前言　/Ⅸ

1　社会科学中的“科学”　/1

　1.1　导言　/1

　1.2　研究设计的主要组成部分　/10

　1.3　本章的主题　/26

2　描述性推论　/31

　2.1　普遍知识与特定事实　/32

　2.2　推论:数据搜集工作的科学目标　/43

　2.3　关于定性研究的公式化模型　/46

　2.4　关于数据搜集的公式化模型　/48

　2.5　概括总结历史细节　/50

　2.6　描述性推论　/52

　2.7　判断描述性推论的标准　/60

3　因果关系与因果推论　/72

　3.1　定义因果关系　/73

　3.2　因果关系的其他定义　/82

　3.3　估计因果效应需要的假设　/88

　3.4　因果推论的判断标准　/93

3.5 建构因果理论的规则 /96

4 确定观测对象 /111

4.1 不明确的研究设计 /114

4.2 随机选择的局限 /120

4.3 样本选择偏差 /124

4.4 有目的地选择观察值 /135

4.5 结论 /144

5 在研究中需要避免的问题 /145

5.1 测量误差 /146

5.2 遗漏变量偏差 /163

5.3 包含不相关的变量会导致估计低效 /176

5.4 内生性 /180

5.5 对解释变量进行赋值 /191

5.6 控制研究环境 /193

5.7 结论 /201

6 增加观察值数量 /202

6.1 基于单样本的因果推论研究设计 /203

6.2 到底多少观察值才算足够? /207

6.3 增加观察值的数量 /211

6.4 结束语 /221

参考文献 /223

1 社会科学中的“科学”

1.1 导言

本书将向读者系统地介绍如何在社会科学领域开展研究。我们本着实用目的撰写本书：主要向读者介绍研究设计的方法，使之能够被应用于社会及政治生活并获得有效推论。虽然我们将重点放在政治学上，但这本书的内容将同样适用于社会学、人类学、历史学、经济学、心理学等其他学科；同时本书提供的方法也可以应用在某些非学术类的专业领域，比如司法证据、教育研究和临床推理等。

本书既不是关于社会科学一般原理的论述，也不是针对问卷设计、田野调查或统计数据分析等具体技术的操作指导。确切地说，这本书将要告诉读者如何设计研究方案，即怎样提出研究问题、开展学术研究并进行有效的描述及因果推论。因此，本书介于抽象的哲学论述和实用的研究方法之间，重点是论述社会科学研究中蕴含的基本逻辑。

1.1.1 基于同样推论逻辑的两种研究风格

我们的目标是通过一个共通的推论逻辑将通常所说的“定性研究”和“定量研究”联系在一起。这两种研究模式的差异是显而易见的，某些时候甚至会相互冲突。但在我们看来，它们之间的差异仅仅体现在研究风格及具体研究方法的差别上。其实，这两种研究的框架的逻辑

都是相同的。在针对定量研究的讨论中,我们将详细阐述这一逻辑并使之规范化和公式化。当然,这一推论逻辑也存在于那些高质量的定性研究中。总之,不管是定性研究还是定量研究,在研究设计中准确把握该逻辑将大有裨益。

众所周知,定性研究和定量研究的“风格”迥异。定量研究侧重计算和统计学分析,其研究建立在对一些现象特定方面的数量化测量上。这种方法从个别现象出发,将个别现象加以概念化以便归纳出一般性的描述或以此检验因果假设;同时,该方法一般使用那些能被其他研究者复制的测量和分析技术。

相对而言,定性研究所使用的方法往往没有定式。不过显而易见的是,定性研究所使用的方法大都不依赖于数量化测量。采用该方法的研究者通常集中在一个或若干个案例上,通过集中访谈或对历史资料的深入分析开展研究。定性研究倾向于使用推论方法,注重对事件或个体进行全面和综合的分析。尽管依靠的案例有限,定性研究依然可以从这些有限的资源中挖掘出大量信息。在社会科学中,定性研究往往与实地调研或个案研究有相同的地方:都侧重于特定案例、决策、制度、定位、议题或者某项立法。在定性研究中,对于案例的选择取决于它们的重要程度,在这点上,两种研究风格其实是一致的。例如,某个国家发生的重大变革、选举、重要决策或世界性的危机;又比如为什么民主德国政权在1989年会突然分崩离析?更概括地讲,为什么几乎所有的东欧共产主义政权都在1989年瓦解了?有时候研究者也从一系列特定事件中选取一个作为案例,比如一场政治革命或某个社区居民反对在其周围建垃圾处理场。而有时候定性研究者也以实地研究的方式开展研究,比如研究某一特定历史与文化状况。一般来讲,对这些特定地区或事件的研究一般都需要深入的分析与详尽的考察。

几十年来,关于案例研究和统计研究、实地研究和比较研究,以及“科学”的定量政治科学研究和基于大量文献的历史研究之间孰优孰劣,政治学家们始终争论不休。一些定量研究者认为,系统的统计学分析才是通往社会科学真理的唯一道路。而对这一观点,定性研究的支持者却抱着强烈反对的态度。激烈的争论由此产生并持续至今。很不

幸，对特定研究风格的偏向导致了社会科学相应地被分成两大流派：定量—系统—归纳派与定性—人文—推论派。前者所依赖的统计分析越来越复杂（令那些没有相关专业背景的人迷惑不解），而后者则渐渐痴迷于分析那些令人感兴趣但不可复制或无法归纳的事件。

本书的一个主要目的是试图说明定量研究和定性研究之间的差别仅仅是风格和方法论上的不同。其实，这种差别在实质上也是不重要的：无论采用定量方法还是定性方法，所有高质量的研究所依赖的逻辑都是相同的。基于这两种方法的研究都是系统和科学的工作。比如，在历史研究中通过有效的因果推论对所有可能的解释进行评估就是分析性的。从这个意义上说，历史学（或历史社会学）与社会科学之间是互通的（Skocpol，1984：374—386）。

为了清除上述理解上的障碍，我们首先需要廓清“定性研究”的概念——虽然本书的主题是定性研究，但这并不意味“定性研究”和“定量研究”在风格差异之外还存在本质的区别。

实际上，大部分研究工作都不能被简单地划归到其中一类，好的研究方案总是试图将两种方法加以综合。比如，在同一个研究中，需要收集某些数据用作统计分析，但对于一些重要信息，比如人们的思想及领导风格上存在差异，却一般不采用实证方法处理。定量研究更适用于对社会、政治或经济行为模式与趋势的分析。为了把握瞬息万变的社会，我们需要学会概括那些不容易被量化的信息。此外，所有的社会科学研究项目都离不开比较分析，这需要我们判断事物是在程度上还是在性质上更为相似。

我们以最近两个优秀的研究为例来说明上述观点。在《强制合作》（*Coercive Cooperation*，Lisa L.Martin，1992）中，作者通过定量分析第二次世界大战以后发生的 99 项经济制裁案例，来探讨国际合作在经济制裁中发挥的作用。虽然这些分析提供了大量有价值的信息，但仅由这些信息去获得因果推论还是有些含混不清。为了克服该问题，作者在所有样本中着重分析了六个案例。在另外一本著作《使民主运转起来》（*Making Democracy Work*，Robert D.Putnam，1993）中，帕特南（Putnam）和他的合作者对意大利地方议员进行了多次访谈：1970 年

受访议员人数为 112 人,1976 年为 194 人,1981 年至 1982 年则增至 234 人。同时,他们还在 1976 年访谈了 115 名地方领袖;在 1981 年至 1982 年,受访的地方领袖人数为 118 人。在 1983 年他们又通过邮寄问卷的方式将该调研范围扩展至全国,使之进一步涵盖了 500 多名地方领袖。在 1977 年至 1988 年期间,四次针对此研究的全国性普查也因此展开。此外,在 1976 年至 1989 年间,该书的作者们选取了意大利的六个行政区,就地方政治的发展问题进行了详细的案例分析。调研期间,通过与政治活动家密切接触,帕特南等人"对意大利政坛中决定过去 20 年来地方政治发展的政治运行规则及决策者有了深入的了解"(Putnam, 1993:190)。

从这些例子中可以很清楚地看出:无论针对何种研究问题,定性研究和定量研究之间都各有优势并相互补充。我们不鼓励完全采用定量方法,因为通过对量化数据进行统计学假设检验的方式并非适用于所有社会科学研究对象。当然我们的意思也不是让所有的社会科学研究者从图书馆迁至计算机室,或用程式化访谈取代有针对性的对话。确切地说,本书试图告诉读者的是:如果定性研究者能够关注那些在定量研究中已经普及的科学推论原则,他们所从事的非统计研究的结论将会更加可靠。从严格意义上来说,那些定量研究基于统计学的建模方法对所有研究都适用,甚至对那些没法用定量测量的变量也有效。这是因为使用统计模型能够让推论逻辑更加清楚。

当然,本书所说的推论逻辑并非适用于社会研究者关注的所有问题。许多政治生活中的重大问题,如机构、义务、合法性、公民权、主权的界定,以及国民社会与国际政治之间关系的定位,等等,更多是隶属于哲学而非实证范畴。尽管如此,这一推论原则依然适用于所有以探索世界为目的的研究。实际上,社会科学之所以有别于通常意义上的观察,原因就在于它的目的是通过规范性地使用研究程序来获取有效的推论。本书集中在实证研究,这意味着我们将回避社会科学研究中涉及的哲学问题,比如关于后现代主义的争议、真理的本质及存在、相对论等。当然,我们承认世界中可能存在一些尚不为人知的科学,但毕竟它们还未有定论。

此外，本书所强调的推论原则并不意味着在做出有效推论前必须进行万无一失的实验或收集到所有数据。对一项重要的问题，即使研究者手头的信息很少，也有值得开展研究的价值。在这种情况下，无论采用哪种研究方案都可能得到不太确定的结论。只要我们如实叙述出这种不确定性，该项研究依然很有价值。由于社会变化太快，信息不足在研究过程中难以避免，所以有时候获得的结论的确定性不高。不过，研究者仍然需要对社会现象进行描述并尽可能地解释。有些时候，待研究的问题具有时效性，以至于条件不允许通过最有效的科学方法来收集数据。就像一个精神错乱的人已经在我们面前抡起斧子，让他先做一份五页纸的精神病测量问卷显然不是明智的选择。

正如约瑟夫·熊彼特(Joseph Schumpeter)曾经引用阿尔伯特·爱因斯坦(Albert Einstein)的评述：“如果要论证确凿无疑，它们就一定与现实无关，只要论证涉及现实，就必然具有一定的不确定性。”(Schumpeter, [1936] 1991:298—299)虽然完全的确定性难以实现，研究者仍然可以遵从科学推论的方法来提高研究的信度、效度、确定程度及真实程度。社会科学的目的就在于对世界作出描述性和因果性的探索。对那些不能容忍片面及非完美假设，或不以描述及因果推论为目的的研究，就只能采用其他方式去获得灵感了。

综上所述，本书的目的不是为科学的实证研究提供诀窍。我们在书中试图揭示的一系列规律和原则，其目的在于规范而非扼杀研究思维。无论定量研究还是定性研究，研究者所依赖的理论、制定的研究设计及搜集的数据都是不完美的。任何推论准则都有例外情况，但通过这本书的学习，读者可以清楚地界定出这些例外，进而评估其对研究可靠性的影响，最终确定其是否影响结论的不确定性程度。本书所提供的内容当然不是教条，而是训练有素的思维方式。

1.1.2 社会科学中关于“科学研究”的定义

本书所说的“科学研究”是一种理想状态。即使是最严谨的定量研究或定性研究，也只是向完美状态趋近而已。尽管如此，读者依然需要

知道什么样的研究才是好的研究。[1]这种说法对定性研究者而言可能会产生误导，因此我们在下面给出明确定义。需要再次说明的是，定量研究和定性研究之间的差异并不是谁更科学。好的研究，即科学的研究，在形式上既可以是定性的也可以是定量的。不过在具体的方案设计上，科学的研究必须满足以下四个条件。

1. 以推论为研究目的。科学的研究设计建立在关于真实世界的实证信息之上，以获得描述性或因果性推论。在科学的研究中，对特定现象进行详细的描述是不可或缺的，事实的简单罗列和累计远远不够。无论定性研究还是定量研究，信息收集得越系统效果就越好。所谓科学的研究还对这些信息的处理提出了进一步要求：试图从这些直观数据中推导出难以直接观察到的结论。因此，这就涉及描述性推论：利用观察值来研究其他无法观察到的事物；同时也涉及因果推论：通过观测数据来研究其因果效应。推论的范围可以在时间和空间上加以限制，如对美国1960年以来选举的研究、1989年以后的东欧社会运动等。甚至可以设置得更加长远，比如对农业出现以后人类行为的研究。在上述两种情况下，科学研究区别于其他工作的地方就是超越现有的观察范围从而获取推论。

2. 采用公开的研究程序。科学的研究采用明确、系统且被同行公认的研究方法进行数据的收集及分析工作，以确保其有效性能够被评估。在这点上，大多数定性研究在研究程序及推论上很少有明确的标准可循。正如罗伯特·默顿（Robert K.Merton，[1949]1968：71—72）指出："社会科学对定性数据的分析与理解往往禁锢在狭隘世界里，晦涩而难以言喻。尽管如此，科学的方法应该是公开的，而不能独属于某一个体或者群体。"虽然默顿的评论并非针对所有的定性研究，但这些评论对某些定量研究同样一针见血。不过很多定性分析看起来确实如他所说，甚至流露出如果存在明文规定的原则就会抑制其创造力一样。

① 在本书中我们不采用"类似实验"（quasi-experiment）这样的说法。在研究设计中，如果研究者能够做到有目的地选取样本及控制关键解释变量的取值，这样的研究就是实验；如果不能做到，就是非实验研究。当然，无论实验还是非实验研究，都有相应的优缺点。

他们通过某种方法观察现象,提出问题,然后根据这些样本获得更多信息并最终得出因果推论。如果没有明确说明观察推论的方法与逻辑的话,其他学者便无法判断其结果的有效性。一旦无法对样本选择、处理标准以及获得结论的逻辑做出评价,读者就不可能从这样的研究中学到什么东西,可重复性更是无从谈起。这样的研究并不是一个公共行为。也许它有一定的可读性,但不能视为对社会科学的贡献。

3. 结论是不确定的。很显然,推论过程往往是不完美的。推论的目的是用定量和定性的数据去了解周围的世界。从不确定的数据中妄图得到完美、确定的结论是徒劳的。因此,不确定性是所有研究与认知中不可或缺的方面。没有对不确定性进行合理的估计,对真实世界的表述以及因果推论就会变得含混不清。如果一个研究者刻意回避不确定性,他/她要么通晓万事,要么根本不知道其结果是否具有确定性。无论如何,缺乏不确定性的推论并不是我们所说的科学。

4. 科学研究的内容是关于方法的。科学研究遵守一系列能够确保推论有效性的规则。这本书的主要任务正是解释这些重要规则。①“科学”的内容并不是研究的题材,而是研究的方法和规则。因为我们以此来研究这个世界。关于这点,一个多世纪前,卡尔·皮尔逊(Karl Pearson, 1982:16)就解释道:“科学的领域是没有界限的;研究可用的材料也是无穷无尽的;每一个自然现象、每一段社会生活、过去和现在发展的每一个阶段都可以被当作研究的材料。”能把所有科学研究整合起来的是其采用的方法,而不是那些用于研究的材料。

以上关于科学的四个特点还有进一步的含义:科学从某种意义上来说很像社会中的公司。每一个研究者或者团队在既定的知识和领域内进行分工。错误有可能被同行发现,因此是无法避免的。理解科学研究的这些社会特征可以使心态变得平衡,这意味着我们的研究并非一定要经得起考验从而对社会做出贡献。不管是表述问题还是提出概念,不管是理论的建构还是对理论的评价,这些方面都有可能遭到别人

① 本书的内容涵盖了科学推论中的绝大部分重要规则,但并没有包含所有。事实上,大部分哲学家都认为他们做不到完全彻底地归纳推理,即使在原则上也无法做到。

的质疑。但只要我们的研究明确地涉及(或者是试图转变)学者们关注的事物,并采用了公开的科学方法,得出了与科学规则及相关信息一致的推论,就可以视作贡献。在这个意义上,一篇小文章的贡献也胜过那些永远放在抽屉里或者保存在计算机硬盘中的“大作”。

1.1.3 科学与复杂性

社会科学试图弄清那些我们认为或简单或复杂的社会状况。不过必须承认的是,复杂性并不完全是一成不变的:这个世界当然不会自然地被分为简单和复杂两类事物。与此相反,认识上的复杂性部分取决于研究者如何对现实进行简化。而简化能力取决于是否能够以连贯而有逻辑的方式说明结果和解释变量。在这一过程中,更多的观察会有帮助,但依然不够。因此,“复杂性”部分地取决于我们对理论的掌握情况。

不管研究复杂还是简单的事件,采用科学的方法是非常重要的。复杂性可能使推论具有不确定性,但不会使它们变得不科学。不确定性与数据局限性不应该是导致我们放弃科学方法的原因。与此相反:遵循科学推论原则的最大好处正体现在当数据受限,测量工具有缺陷,测量结果模糊以及观察不确定的境况下。如果数据是明确、清晰的,方法可能就不太重要了。因为即使借助不那么完美的推论规则,我们也能获得大致正确的答案。

让我们考虑一些体系复杂并且独特的事件。罗马帝国的崩溃、法国大革命、美国第一次内战、纳粹大屠杀以及 1990 年两德统一都是这样的例子。这些事件产生于许多因素的交叉作用,众多因素的交汇对于事情的发生起着重要的决定作用。也就是说,各种事件和因素发生的先后在特定时间、地点的汇合而产生的交叉作用使得观察到的这些事件得以发生(Hirschman, 1970)。此外,人们也很难相信这类事件是大型历史因素的必然产物:其中有一些取决于个人的特质、机制或社会运动。事实上,从理论的角度来看,机会同样发挥了作用:理论范畴之外的因素在事件的发展过程中为其发生的顺序提供了至关重要的连接作用。

对这些事件进行概括是了解它们的方法之一:把每个案例都概念化为一类事件中的一个组成单位并对其进行有意义的概括。这种方法对于分析一般的战争或革命往往行之有效。不过也要注意那些属于特例的战争和革命,它们构成了统计分布中的“异常值”。此外同样需要引起注意的是,早期著名的革命和战争可能对随后发生的同类事件有显著的影响,比如法国大革命。研究者在比较它与后续事件的时候就必须十分谨慎。因为后续类似事件一定程度上可能是对其的模仿。增加该类事件的样本可能会起一定作用,但也不是在所有情况下都合适。

另一种处理这些不常见的、大型事件的科学方法是进行反事实分析:“对事件过程的理解,会因一个或多个条件的改变而变化。”(Weber, [1905] 1949:173)对这一观点系统化、科学化的应用,可以通过以自然科学为导向的地质学与进化生物学中关于偶然事件的极端个案来诠释。斯蒂芬·古尔德(Stephen J.Gould)曾经提出,从随机的、偶然的事件中区分关于进化发展的系统化特点的方法之一是去想象。如果某一特定时点上所有的条件都固定不变,历史继续发展,那么这个世界将会怎样?他认为,如果能“重播生命的磁带”,让进化再次从头开始,那么今天世界上的生物将是完全不同的(Gould, 1989a)。

研究生物进化的学者最近关注的事件是恐龙为何在6 500万年前突然灭绝。按照古尔德(1989a:318)的说法:“我们必须假设如果天体碰撞没有殃及恐龙,那么意识就不会渐渐地进化和发展出来。”如果这一说法是正确的,恐龙灭绝将与任何人类历史中的重大事件同样重要;然而,恐龙灭绝不能运用简单的一般规律进行系统的、比较的研究。

虽然如此,恐龙灭绝仍可以被科学地加以分析:其替代假说可以预测后续现象并由此反过来检验假设。20世纪70年代末(W.Alvarez & Asaro, 1990),路易斯·阿尔瓦雷斯(Luis Alvarez)和他的合作者在加州大学伯克利分校提出了一个解释恐龙灭绝的假说——天体碰撞假说:一颗陨石以大约每小时7.2万千米的速度撞上地球,产生的爆炸当量超过全面核战争。如果这一假设正确,它将意味着在6 500万年前的地层中应该含有一种陨石中常见但在地球上十分罕见的元素;事实上,在相应地层发现铱元素为该理论提供了部分有力的证据。毋庸置疑,天体碰撞

是一个特殊事件，但应该存在一些可被观察到的现象①：比如，可能在地球的某些地方发现陨石坑（有几位研究者声称已经找到）。②

恐龙灭绝之谜仍然没有被完全破解，不过围绕该问题的争议已经催生出许多有价值的研究。举这个例子是想说明，科学的概括对于理解那些极不寻常的偶然事件是很有用的。阿尔瓦雷斯的假说虽然不能通过一些普通事件得到验证，但它仍然可以被某些后续现象加以检验。然而，读者应当注意到，只有得到经验评估并且通过一系列严格的验证，假设才能变成合理的解释。合理解释的最低要求是它的预测现象必须符合大家关于外部世界的认识；在最好的情况下，它能够预测到被伊姆雷·拉卡托斯（Lakatos，1970）称为"新事实"的东西，即那些以前没有观察到的事物。

问题的关键在于研究者需要提高理论、数据和对数据的利用效率。只有这样，诸如恐龙灭绝之类的独特事件就可以被科学地加以研究。通过澄清概念和规范变量来改善理论，可以获得更多可观察的现象，甚至可以检验那些关于特殊事件（如恐龙灭绝）的因果理论。改善数据质量可以帮助研究者观测到更多的现象以及提高对数据的利用效率，进而从现有数据中获取更多的信息。哪怕研究对象非常复杂，谨慎的研究设计也是必不可少的。如果研究者都尽可能多地收集那些与预测现象相关的数据，不管研究多少现象，哪怕只有一个，研究质量也可以得到改善。

1.2 研究设计的主要组成部分

社会科学所追求的是在一个完善的科学探究结构中进行洞察和发

① "可被观察到的现象"（observable implication）是本书中经常出现的概念。该概念指的是如果理论是正确的，研究者就会相应地在真实世界中观察到一些与该理论预期一致的现象。——译者注

② 然而，另一种认为灭绝是由于火山爆发的假说也被铱元素的存在所支持。该假设与天体碰撞假说相比较，其预测的现象似乎更加符合认识：并非所有物种同时灭绝。

现的创造性过程。一流的社会科学家一般都不同意研究设计是一个关于机械性收集数据及评估过程的计划。相反,学者必须具备灵活的思维,能够推翻观察世界的传统方式,进而提出新问题以及适当地修改研究设计,然后更多地收集比预期更丰富的数据。如果研究人员的调查结果是有效的并能被该领域的其他学者接受,那么所有这些修改及反思都必须按照明确且符合推论规则的程序进行。总之,调查过程应该在一系列固定的规则下展开。

社会科学家往往从一个设计好的计划开始,然后收集数据并从中获得结论。不过,这一进程并不总是一帆风顺。此外,即便按这样的顺序进行的研究也不能保证做到最好:因为并不总是简单依照研究计划及数据收集就可以获得结论。研究者依据研究设计去收集数据,但他们常常发现主要的研究问题、理论与已获得的数据之间并不契合。在这个阶段,研究者会因此灰心丧气,错误地认为只有其他学者才能找到与研究最密切且直接契合的数据,而唯独他不能。产生这种悲观情绪的原因在于研究者在完成他们的智慧建筑物后,拆除掉脚手架,却没有留下一点痛苦和不确定性的痕迹。因此,这个研究过程似乎比实际情况更加机械和呆板。

我们对那些试图将理论联系数据的研究者提出一些建议。有时,他们可以设计更合适的数据收集方法以更好地来检验理论;也可以利用已有数据重塑一个理论问题,并形成一个更重要的研究项目(甚至可能提出一个未知的、全新的问题)。如果新研究遵循推理的规则,它就仍然是科学的,并能获得可靠的结论。

研究者应尽量在实际操作前完善研究设计。当然,数据可以约束思想却不完全受思想约束。一些研究在收集完数据后就停滞不前,这样的现象并不少见。这并不意味着理论是错的,而是搜集到的数据不适合回答最初提出的那些研究问题。因此,从一开始就应该试图去了解哪些能做,哪些不能做。这可以使得研究者在着手研究设计的时候,就能预测后续可能碰到的问题。

为了便于分析,本书把研究设计分为四个部分:研究问题、理论、数据和数据的使用。这四个部分并不互相独立,学者一般也不会以任何

既定的优先顺序来处理他们。事实上，对那些还没有确定主题就开始实际工作的定性学者来说，数据往往是最先需要考虑的，其次才是其他部分。本书将在1.2.1—1.2.4小节中详细解释该分类对理解研究设计性质的特别用处。为了了解资源被重新定向后我们究竟可以做些什么，本节余下的部分都假定研究者拥有无限的时间及资源。当然在实际情况下我们必须为此做出妥协。理解接下来本书对这四个部分的建议能够使做出的妥协最大限度地改善其研究设计。虽然事实上，研究显而易见地会受到外部因素的制约。

1.2.1 改进研究问题的质量

在研究目标确定之后，研究者就应该思考接下来该做些什么。那么对一个研究问题，以何种方式操作才能获得关于社会和政治现象的有效解释呢？我们首先从研究问题讨论，然后再考虑各阶段的设计及具体步骤。不过，研究问题从何而来？学者如何去选择主题？针对这些问题并没有简单的答案。同其他人一样，卡尔·波普尔（Karl Popper, 1968:32）认为："并不存在一个逻辑的方法来帮助产生新思想……发现新思想的过程包含着'非理性因素'。"在研究最初阶段涉及的规则，并没有某些研究议题中涉及的规则那样规范。比如，关于社会选择中涉及的实验设计可以依靠现有文献，公众对公共政策态度调查的抽样也有统计标准可依，对科层制进行观察也可以按照相关指南操作。但如何选择研究项目却没有规则可循，是否进行实地工作、在哪儿进行同样没有一成不变的规则。

一般来说，研究者可以采用在社区中抽样来研究替代教育政策的影响，或对种族冲突的方式进行概念化以便制定和测试其发生概率。但没有现成的规则告诉我们，是否需要研究教育政策或种族冲突。有时，利用社会科学的方法可以区分出好或不好的研究。关于民主德国政府在1989年解体的研究方法就有好有坏。关于研究候选人在税收问题的立场对成功获选可能性的影响所采用的方法同样有优有劣。但没有什么方法可以区分出是研究民主德国政权崩溃更好，还是研究税

收在美国选举中的作用更有价值。

社会科学家研究的具体问题也可能与个人兴趣有关。某一群体中的研究者对其所在的群体进行研究是相当常见的:妇女、黑人和移民就常常在他们各自的研究历史上引路。个人倾向和价值观也可能影响到话题的选择。与研究国家政策制定的学者相比,研究第三世界国家政治的学者往往更倾向于旅游,并且更能忍耐艰苦的生活环境;而研究国际合作的学者则可能特别讨厌暴力冲突。

个人经历和价值观可以成就一个人成为社会科学家的梦想,并在其后促使他们选择一个特定的研究问题。就此而言,研究人员可能会很恰当地构想出他/她从事某一研究项目的“真正”理由。不过,无论选择主题的原因多么个人化、特殊化,本书所讨论的关于科学推论的方法及原则都将有助于学者制定出更为有效的研究设计。从对社会科学的潜在贡献来看,出于个人原因来选择研究主题既不必要也不足够。在大多数情况下,个人因素不应该出现在我们的学术著作中。直截了当,甚至是不那么礼貌地来说,没人会在意你想的是什么——学术界只在乎你能证明什么。

虽然没有关于如何选择主题的明确规则,研究者仍然能找到办法(个人喜好除外)向学术界证明研究的可能价值。在理想状况下,所有社会科学研究项目都应该满足两个标准。第一,研究项目应该提出对现实世界真正重要的问题。该问题应该首先是政治、社会、经济生活的产物。同时,该问题能够有助于探究哪些因素显著影响了人们的生活,了解和预测那些可能有害或有益的事件(Shively, 1990:15)。第二,一个研究项目应该通过提高研究者作出科学解释能力的方式对文献做出具体贡献。后一个标准并不意味着所有为社会科学解释添砖加瓦的研究必须获得因果推论。有时候在着手解释问题之前,某个领域的知识更多需要的是事实调查和描述性说明。有时候某个项目的贡献就在于它的描述性推论。甚至有时候研究目标都谈不上描述性推论,只是对事件近距离的观察和对历史细节的粗略描述。由于这些工作是解释问题的先决条件,因此也符合上述第二个标准。

第一个标准指导研究者关注现实世界的政治及社会现象,关注当

前和历史上那些塑造人们生活的事件。一个研究问题是否符合这个标准从本质上说是一种社会判断。第二个标准将关注点引向社会科学的学术文献中那些尚未提出的但极具意义的问题、那些已经提出但是没有被解决的难题以及用来解决这些问题的科学理论与方法。

政治科学家很容易发现符合第一个标准的主题。过去四百多年的十场主要战争已导致近3 000万人死亡(Levy, 1985:372);一些"有限度战争",比如美国的越战以及美国与两伊(伊拉克与伊朗)的战争,均导致超过100万人口的死亡;一旦发生全面核战争,将会造成数十亿人的死亡。内政与外交上的管理不善导致经济萧条在全球蔓延,正如20世纪30年代区域性与地方性的经济大萧条。20世纪80年代许多非洲和拉丁美洲国家的悲惨经历也证明了这一点。跨国政治体制的变革与普通民众生活状况的巨大变化是紧密联系在一起的。这反映在即使经济发展水平差不多的国家,它们在婴儿死亡率和预期寿命等福利指标上也存在着巨大差异(Russett, 1978:913—928)。在美国,那些旨在减轻贫困、防止社会分化的项目,功效亦差别很大。毫无疑问,哪怕对于理解上述某个议题有很小的贡献,相关的研究工作也都是很重要的。

虽然有很多重大问题可供社会科学家研究,但用于了解这些问题的工具往往是有限且粗糙的。大量关于战争和社会阴暗面的作品对理解这些问题其实并没多大帮助,因为它们既没有系统化的描述现象,也没有提供有效的因果性或描述性推论。精彩的见解当然有助于产生有趣的新假设,但才华本身并不是一种实证的研究方法。只有经得起实证的检验,假说才能真正做出贡献。从这个意义上来说,本书的内容并不能助你成为一个才华横溢的人。它强调的是进行研究的重要性,以此对现有的知识体系做出贡献。

我们选择研究问题的第二个标准:做出贡献。该标准是指在现有社会科学文献框架内明确地定位出一个研究设计,以此确保研究者了解"当前研究的最前沿",并尽可能减少与现有研究的重复。此外,它还能保障所做的研究对其他研究者来讲也是非常重要的,进而提高整个学术界的研究。可以通过很多方式对现有文献做出贡献。在这里我们列出如下几种可能的方式。

1. 选择学者在文献中十分重视但是此前没有人做过系统研究的假设。如果能找到支持或证伪这一假设的证据，这将被视为一项贡献。

2. 选择一个虽然在文献中被公认但仍然值得怀疑的假说（或一个尚未得到充分确认的假说），以此来探讨这一假说是否错误或者找到其他可供替代的正确理论。

3. 试图解决现有学术界关于某一问题的争论或为争论的一方提供进一步证据，也许可以证明该争论从一开始就没有事实根据。

4. 对研究进行设计以阐明或评估文献中那些无可置疑的假说。

5. 找到并确认一个在已有文献中被忽视的重要问题，然后进行系统性研究，由此对该领域做出贡献。

6. 证明一个文献中为某些目的而设计的理论或证据可以用来解决另一个文献中存在的问题。

过分强调对学术文献作贡献而忽视那些有实际价值的议题，可能会使我们的研究变成在政治上无意义的问题。相反，只关注当前的政治热点而不考虑一个问题是否符合社会科学知识框架的研究也仅会徒增与知识积累无关的工作。

这两个如何选择研究问题的标准并不是对立的。从长远看，使用科学方法产生和检验解释性假设会增强我们对现实世界众多现象的理解。从短期来看，实际用处和长远科学价值之间有时会相互矛盾。例如，曼昆（Mankiw，1990）就指出关于宏观经济的理论和应用在20世纪七八十年代以后经历了急剧的分化：那些已经被实践证明在理论上不符合逻辑的模式仍然被用来预测美国经济的走向；那些旨在纠正这些缺陷的新理论模型仍然在逐步完善的过程中，同样不能期望它们做出准确的预测。

当一个研究者选择主题时，现实世界中的实用性和为科学进步作贡献这两个标准有时确实是互相对立的。一些研究者可能从一个有重大社会意义的现实问题出发，比如核战争的威胁、男女收入差距和东欧国家的民主转型等。其他研究者也可能从一个源自社会科学文献的问题着手，比如针对在不确定条件或信息不对称状况下个体决策的若干次实验，其结果可能相互矛盾；采用国会选举理论作出的预测与最近的

选举结果并不一致。对这两个标准的区分并不是硬性的。一些研究问题从一开始就能满足上述两个标准，只是其研究设计往往更加接近其中一个。①

不管从哪里着手，旨在解决问题而设计的研究都应该尽量满足上述两个标准。显然，我们前进的方向取决于从哪里开始。如果需要解决社会科学中的难题，那就必须考虑如何使研究课题与那些具有重要意义的现实问题相契合。举一个例子，政治决策者如何利用实验方法来指导现实世界中的战略选择，或去预测某一理论可能会导致的行为后果。如果从现实问题着手，研究者就应该考虑怎样应用现代科学方法展开研究，使其有助于推动现有文献。我们认为，从一个标准移动得太远并非最有效的方法。那些运用实验方法的学者就认为寻求外在参照(external reference)还言之尚早，优化控制实验环境的理论和方法还能取得明显进展。在长期的研究项目中这可能是正确的，但旨在解决现实问题的学者可能认为，在作出解释前必须有准确的描述。该观点可能是正确的，因为在解释性研究计划中，准确描述是一个重要的步骤。

在任何情况下，研究项目甚至具体的研究方案都应该尽可能满足这两个标准：选择一个重要的现实主题，直接或间接地对学术文献做出贡献。本书关注的目标是使定性研究更加科学化，因此我们将更多地讨论那些从真实世界角度出发的研究。不过书中的分析与两个标准均相关。

如果不是从现有文献出发，而是从一个重大的现实问题出发，研究者就必须设计一个可操作的计划。如果一个主题不能提炼成可以进行有效描述或因果推论的具体研究计划，研究者就应该在研究过程中反复修改以达到目标，否则就干脆放弃。一个不能做出学术贡献的拟议

① 社会科学界面临的这个情况和自然科学中关于进行应用性研究还是基础性研究的争论相似。短期来看，针对特定药物或者疾病的应用性研究可以提供医疗服务水平，但对生物机制领域中基础性知识的贡献却不大。基础性研究的角色正与此相反。但多数研究者都认同基础研究最终能够产生强大的应用效果。当然，最好的研究设计应该兼顾两者：既能解决实际问题，又能深化现有基础性研究的文献。

主题同样应该经过不断的修改。一旦基于常识而选定了一个研究主题后，我们就需要与文献展开对话。哪些感兴趣的问题已经被现有文献解决了？我们应该如何提出并完善研究问题以使它能被现有研究工具解答？当然也有人从某个热点问题展开研究，这同样必须从社会科学文献和推论中的问题两个方面着手处理。

1.2.2 提高理论的质量

社会科学理论是对研究问题的答案做出一种合理且准确的猜测，同时还要提供可以证明答案正确性的相关证据。理论往往意味着特定的描述性或因果性假设。就一个研究问题来说，理论必须与已获得的证据一致。“理论不能忽视证据。就像试图颁布法规要求广告必须诚信一样，忽视证据的‘理论’不能被称为理论。”(Lieberson，1992:4；同时参见 Woods & Walton，1982)

谋求理论的发展通常被视作研究的开始。理论的发展也许来源于实践，当然并不是一向如此。如果要发展理论，研究者需要事先完成相关的文献综述和数据收集工作，否则研究问题无从提出。在数据收集之外，还有一些常用方法可以评价和改进理论。这里先简要介绍以下几种方法，在本书后面的章节中将对这些方法做更进一步的详细讨论。

第一，选择那些可能错误的理论。在实际中，从错误理论中学到的教训比从那些宽泛的、以至于原则上都不可能出错的理论中学到的要多得多。[①]研究者需要明确回答以下问题：哪些证据能够被用来说明自己的观点是错的？[②] 如果无法回答这个问题，我们就不能得到一个理论。

第二，选择一个可以产生较多预期现象的理论，以此保证该理论可以被证伪。这样的选择策略允许我们使用更多种类和更大数量的数据来检验理论，从而加大理论被证伪的概率。此外，该策略也能保证搜集

---

① 这就是证伪(falsifiability)原则(Popper，1968)。虽然不同的科学哲学对该问题持有不同的立场，但是大多数学者同意这样的原则：理论只有被清晰地描述出来之后，才符合被证伪的条件。

② 这也许是我们系及其他学术部门面试时最常问到的问题。

到尽可能多的证据去支持理论。

第三，理论要尽可能具体。含糊的理论和假说除了令人迷惑外没有任何实际意义。那些被精确陈述且能做出具体预测的理论，更容易被证伪，因此也是更好的。

许多研究者都建议遵循“简约”(parsimony)原则。不过由于该词在平时闲谈以及学术文献中被过多地使用过，以至于作为原则本身却变得有点模糊不清(相关的详尽讨论参见 Sober, 1988)。对简约原则最清晰的定义来自杰弗里斯(Jeffreys, 1961:47):“简单的理论具有较高的先验概率。”①简约因此成为针对世界本质的判断，甚至是假设:世界被想当然地假设成简单的。只有研究者认为这一假设具有较高的确定性时，才能运用简约原则来选择理论。值得一提的是，该原则在物理学中较为适用，不过在生物学中却被认为是荒谬的。在社会科学中，虽然也有人在自己的领域中努力捍卫该原则(比如 Zellner, 1984)，不过大多数情况下，该原则并不合适。简约就其定义来讲实为对世界的假设，因此它决不能作为设计理论的一般原则。但如果研究者对研究对象的简单性程度有所了解，该原则还是能发挥一定作用的。

研究者需要记住，只有非常了解研究对象才能运用简约原则。因此我们不建议研究者把简约作为一个基础性的原则。研究者甚至都不能运用它去回避那些过于复杂的理论。一般认为，理论的复杂性应该与证据显示出的复杂性相一致。如果没有获得足够的证据去支持理论的复杂性，这将会导致所谓的“不明确的研究设计”(indeterminate research design)的问题(详细讨论见本书第 4.1 节)。不过这仅仅是研究设计过程中出现的问题，和关于这个世界的假设无关。

在收集数据和着手研究前，上述建议都是适用的。一旦收集到了数据，就可以利用这些规则来指导我们修正理论并去收集新的数据，从而获得新理论的新观察现象。当然，这个过程既费钱又耗时，甚至已收集到的数据可能毫无用处。如果理论明显需要改进，但在此时无法收

① 该表述被称为“杰弗里斯—林奇(Jeffreys-Wrinch)简单假设”。该概念类似于“奥卡姆(Occam)剃刀法则”。

集更多数据该怎么办呢？在这种情况下，研究者需要非常谨慎和自制。一个学者特别容易倾向对事实背后的数据提出一个“似是而非”的理论，这样对该理论的真实性毫无帮助。即使理论与数据十分契合，理论仍然有可能是错的，并且无法应用于其他数据。人类往往擅长识别那些具有固定模式的事物，却不擅长辨认那些没有固定模式的存在。（很多人甚至能在随意溅出的墨水污渍中发现模式！）尽量不要对理论做出不符合数据的临时或专门的调整，如果必须调整，就尽可能地循规守矩。①

当研究者完成了数据的收集和分析工作并希望改善理论时，仍然会有不知道如何下手的情况。针对这一问题，我们有两点建议：第一，如果预测是有条件地基于几个变量作出的，如果研究者希望去除某个变量时，是可以这样做的。例如，研究者最初假设那些具有先进社会福利制度的民主国家相互之间不会发生冲突，这一假设也可以被进一步扩展到所有的现代民主国家。这样就可以用更多的案例来评价理论的有效性，同时增加其被证伪的可能性。在接触到数据之后，研究者有可能改变原有理论，使其适用于更多的现象。由于理论经过如此修改会导致其更容易被证伪，因此我们不应该通过将理论可以预测的现象限定于那些已经被观察到并且与理论一致的现象上，以此来“挽救”一个不成熟的理论。

但是，与上述相反的做法也是不恰当的。在观察完数据后，研究者不应该通过增加一个限定条件来保证理论在此限定下仍然有效。假设原有理论是，现代民主国家不互相发动战争的原因是其宪法制度。由于发现了不符合“规则”的例外情况，就将民主制度的前提限定为先进的社会福利制度，这样就保证了观测到的数据符合后一个条件，进而保证了假设的正确性，这一方式显然是不恰当的。再举一个例子：假设原有理论是，仅在严重经济衰退时革命才会爆发，可是观察实际情况后却发现有一个案例不是这样。在这种情况下，增加条件也是不合理的，这样的例子有：繁荣时期永远不会发生革命，除非军事薄弱、政府执行高

① 如果研究者选择一个对现实及文献都十分重要的研究主题，学术界会纠正针对该研究的不足：很多学者会采用其他数据复制研究，以此证实或证伪我们的发现。

压统治、经济产生泡沫以及气候变暖。这种提法只是以一种花哨的并带有误导性的方式说:“我的理论是正确的,某国除外。”因为已经发现该理论对该国不适用,所以这个能证伪现有理论的事实导致我们得不到一般性假设。如果没有搜集到新的数据,我们就无法用证据支持该理论的新版本。

因此,数据观测工作之后对理论的修正要遵循如下几条基本规则:减少对理论的限制(这样它就可以涵盖更多的现象,并且有更多被证伪的机会)。如果没有收集到新的数据来检验修正后的理论,研究者也不应该通过增加限制条件的方法去维持原有理论的有效性。如果收集不到新的数据,研究便会停滞,在这点上没有任何捷径可以渡过难关。有时承认自己是错的也不失为一种好的方式;事实上,否定现有理论对学术文献来说不失为一种价值。基于临时修改的理论获得一些站不住脚的结果而放弃了那些可靠的否定结果显然是一件得不偿失的事情。

即使我们错了,也不必在承认缺陷后就停止研究。为了应对这样的情况,我们可以在书或文章中增加一章或一节来描写现有理论的推测及下一步的实证工作。在这方面仍然还是有相当大的自由空间的。研究者可以提出一些对理论十分重要并能解决问题的附加条件,修正另一个现有理论或提出一个完全不同的理论。在这种情况下,我们虽然没法得到一个具有很高确定性的结论(除非理论一开始就是错的),但可以设计新研究或数据收集方案以判断推断是否正确。对于未来从事该研究的学者来说,这些都是很有价值的参考性意见。

正如上面所讨论的,我们不得不承认社会科学并不严格地全部按照规则运行:对创造性的需要迫使研究者不时需要抛开书本。此外,数据本身可以帮助我们产生新的想法。有时研究者面对数据时会产生灵感,这些灵感会指导他们建构初步的理论。如果研究者能够使自己与其他人相信依照既定计划对理论进行的修正可以在收集数据之前完成,那么这种修正即便很有限,也是值得尝试的。但需要注意的是,在接受新数据检验之前,理论仍然具有很高的不确定性。

从这些规则中我们可以得出一个重要结论,即试点性项目是非常有用的,尤其是那些必须通过访谈或耗费极大成本才能得到数据的研

究项目。初步的数据也许会使研究者改变研究问题或理论，然后重新收集新数据来测试新理论。这样就避免了用同一个数据去产生和检验理论的问题。

1.2.3 提高数据的质量

“数据”是系统收集到的关于世界的信息要素。数据可以是定性的也可以是定量的。有时人们收集数据是为了检验某一理论。数据有可能在确定研究兴趣之前就着手收集。此外，即使搜集数据的最初目的是评价一个具体假设，研究者也很有可能在之后提出一个全新的研究问题。

无论数据收集被用于确定的还是不确定的目的，有些通用的规则是可以用来提高数据质量的。原则上讲，这种改进数据质量的规则可以和本书 1.2.2 小节讲到的改进理论质量的规则区分出来。不过在实践过程中，就像任何理论都需要数据支持一样，任何数据的收集工作也都需要一定的理论指导(Coombs, 1964)。

提高数据质量的首要准则是记录和报告产生数据的过程。没有这些信息，就无法保证使用标准程序分析数据进而得到无偏的推论。只有知道产生数据的过程，我们才能获得有效的描述性推论或因果关系推论。如果想在一次定量的民意调查中记录数据产生的过程，研究者就必须要求知道抽样和提问的确切方法。在关于定性案例的比较研究中，报告出如何选择少数案例进行分析的规则是至关重要的。我们将在本书第 6 章中为定性研究的案例选择方法提供更多的指导性规范。读者需要记住，比选择好方法更重要的是忠实记录和报告对其他人有重要参考价值的使用方法及必要信息。①

① 我们发现许多研究生很害怕因为共享数据和信息，别人就有可能复制他们的结果。他们担心有人会窃取他们辛勤工作的果实，甚至证伪他们的发现。虽然这些担心很正常，但其实大可不必。出版(至少向其他学者派发其研究论文的副本)或共享数据是确保个人贡献被同行认可的最好方式。此外，共享数据会帮助别人沿着你开创的道路继续研究。他们发表研究时对你的致谢也会提高你的知名度和声誉。

在本书1.2.2小节中,我们认为理论可以产生许多可观测的现象。用以提高数据质量的第二个准则就是要收集尽可能多的这种现象,以便更好地检验理论。这就意味着要尽量收集来自不同背景的数据。研究者所观察的关于理论的每一个现象都可以用来检验其真实性。符合该理论的可观测现象越多,解释就越有力,其结果的确定性就越高。

为一个理论的可观测的现象增加新数据时,我们可以通过搜集更多的关于被解释变量(dependent variable)的观察值,或者记录新增的被解释变量。例如把被解释变量分解到更小的时间段或更小的地理区域去研究。此外,还可以收集那些不是很直接相关的被解释变量信息;如果其结果与理论预测的相同,我们对这个理论就会更加有信心。

以理性威慑理论为例:想发动战争的国家需要计算攻击其他国家的成本和收益。这些核算也需要把受到其他国家报复的影响纳入考量范围。对这一理论的直接验证则是为了检验在有战争威胁的情况下进攻与否是取决于攻击者与防卫者间的军事实力权衡,还是取决于防卫者可能获得的利益(Huth, 1988)。如果只使用那些威胁已经发生的情况,这就意味着仅仅收集到部分观察值(单独使用这些观察值可能会导致样本选择偏差问题)。因为那些威胁被制止的情况没有被纳入样本。就这点来说,在测量国家是否有动力做出威胁举动的基础上去收集额外的被解释变量的相关数据(也就是其他的可被观察的现象)也是值得做的。

即使缺乏关于国际政治威胁的数据,我们仍然可以测试那些基于相似假设的其他理论。虽然这是对不同条件下不同被解释变量的研究,但它们都是同一个理论的可观测的现象。例如,研究者可以在模拟条件下进行一个实验,观察"威胁"是被制止还是被军事力量和谈判加剧。也可以考虑是否存在其他类似情况,比如几个寡头公司争夺市场份额或有组织的犯罪家族争夺地盘。在使用威慑战略时,他们如何在不同条件下获得成功。事实上,产业组织领域的经济学家已经使用基于非合作博弈论的威慑理论来研究市场准入和定价策略等问题(Fudenberg & Tirole, 1989)。博弈论认为,由于理论的相似性,支持博弈论假设的经验证据会增加国际政治中国家行为相关理论的合理性。尽

管这一结论在不同领域中的适用性仍然不确定，该问题的重要性也促使我们尽最大努力去思考及搜集相关证据。

无休止地收集数据而不做任何分析只会破坏有效研究的完成。但在实践中，有限的时间和资源始终限制着数据收集工作。虽然更多的信息与案例、额外的调查、其他变量等总能在一定程度上提高推论的确定性，但信息过多与信息过少都能葬送有发展潜力的研究者。坚持再多读一本书或收集更多的数据而不着手写作经常会导致研究者的低产。

第三个准则是最大程度地提高测量效度(validity)。该概念指的是测量到了我们想要测量的事物。失业率可能是衡量国家经济状况的一个很好指标，但两者的含义并不完全相同。总的说来，最大化测量效度最简单的方法就是依靠数据说话，而不让那些不能观察及不能测量的概念妨碍我们。如果被调查者对问题一无所知，研究者应当知道他为什么不明白。只有这样，测量工作才是有效的。如果他真正的意思是表示一个完全不同且也没有把握测量的概念，这样的测量就不具备有效性。比如，在那些政府高压的国家中，很多人回答的“不知道”可能就是一种政治声明的方式；而对其他人来说这只是一种说“我不知道”的方式。两者之间的差别显而易见。

第四个准则是确保收集数据的方法是可靠的。可靠性意味着同样的方式和程序可以产生相同的结果。在测量目标“真实”状态保持不变的前提下，在不同的时间使用同一程序将观察到同样的结果。①不同的研究者采用同一测量程序也应该能得到相同结果，当然这取决于程序的准确性。②

① 研究者可以分两次测量同一对象，观测结果是否相同，以此检验可靠性。在实际中这也比较容易操作，比如在同一次访谈的不同时间段问受访者字面意义相同的问题。但值得注意的是，第一次提出的问题可能会影响受访者第二次的回应，研究者必须确保两次测量彼此独立。

② 关于该点的一个例子是使用几个编码员来提取访谈记录中的系统化信息。如果两个人使用的编码规则相同，我们就可以据此得知他们做出相同判断的频率。如果他们无法做出可靠的测量，就意味着需要更为精确的编码规则，然后重新测试。最终研究者需要制定一个规则，确保不同编码员使用相同的程序获得相同的结果。

最后一个准则是确保所有数据及分析工作可以被复制。复制不仅适用于数据以检验采用的措施是否可靠，也适用于产生结论的整个推理过程。例如，根据我们的研究报告，即使是一个新的研究者，也能够复制我们的数据及推论的逻辑。不管有没有人复制自己的研究，可复制性都是非常重要的。只有报告出足够多的研究细节，才能使得研究被别人复制，由此来评估研究遵循的程序和使用的方法。

在某些研究中，数据往往难以或无法被复制，比如受访者可能死亡或失踪，目击者或参与者无法重现真实事件。复制在不同研究传统中具有不同的含义。定量学者注重对相同数据的重复分析。任何试图复制他人研究，甚至是那些杰出工作中的定量结果的人都深知复制比看起来要难很多，也往往比最初设想的更有价值(对于定量研究中可复制性的详尽讨论见 Dewald et al., 1986)。

脚注和参考文献为传统定性研究提供了类比性。后续学者能够通过这些工具找到著作中使用的资料，并且能够对那些基于这些信息的推论做出评估。至于建立在直接观察基础上的研究，可复制性一般比较困难。学者可以借助其他人的现场记录或者调查录音来检查它是否支持原始调查者得到的结论。由于现场调查的数据大多涉及会话、观感以及其他未被记录的信息，因此根据这些数据对结果进行再分析的工作并不常见。当然学者尝试这样复制，还是可以取得一些进展的，并且这会促使研究者保持完整记录现场信息的好习惯。有时，一个包括数据收集在内的整个课题都有可能被其他学者不断复制。由于不能返回实地，复制工作虽不能尽善尽美，但依然有价值。对定性研究最大规模的复制工作可能是在 1929 年和 1937 年两次针对“米德尔敦”(Middletown)的社会学研究。在初始研究五十多年以后依然出版了大量复制该研究的书籍和文章(见 Caplow, 1983a, 1983b，以及其中的引用文献)。当然不是所有的定性研究都需要如此广泛的复制，该研究项目的复制工作确实是可供学习的典范。

所有研究都应该尽量实现可复制性：学者应该记录那些收集信息与得出推论的确切方法、规则与程序，以便其他研究人员可以重复他们的研究，并获得相同的结论。复制也意味着使用未公开的或私人的记

录的学者应确保后续学者同样可以获得材料；为了自己独享此特权带来的好处而不让他人接触材料将导致复制无法实现，研究质量通常也会出现问题。一般而言，研究者的工作未必一定会被别的学者复制，但我们有责任为他人潜在复制的可能性做好准备。即使自己的工作没有被复制，那些为复制准备的材料也有利于读者了解和评价我们的工作。

1.2.4 更好地利用现有数据

与其试图充分利用现有有缺陷的数据，不如通过收集更新更好的数据解决现有数据中存在的问题。当然，后一种方法并不总是可行。社会科学家常常发现自己的数据有问题，但是同时也无法获得更好的数据。因此，最大限度地利用手中现成的数据就成为权宜之计。

如何更好地利用已有数据不仅是统计方法课教授的主要内容，也是统计推断对社会科学做的主要贡献。在统计推断中应用到的那些准则同样适用于定性研究。这本书的剩余部分将更加充分地讲解这些规则。对于如何更好地使用现有数据，以下内容仅仅提供一个简单的纲要。

第一，尽量利用数据获得“无偏”的，即正确的推论。从统计学角度理解这点就是：不论在定性研究还是定量研究中，从若干个数据库中获取数据并运用相同的方法进行分析以此获得结论。由于数据或程序应用中这样或那样的错误，单独应用某种方法未必能得到正确结果。即便这样，研究者在多次运用中，“无偏的”程序就有可能获得正确的结果了，因为该过程不会使结论产生系统化偏差。

当然，能否实现无偏推论取决于原始数据的收集与其后的使用；正如我们之前指出的那样，最好在数据搜集工作之前就预测可能发生的问题。我们之所以在这里简要地提及，是因为使用这些数据时需要特别小心地分析，以确定在数据收集过程中是否忽视了偏差产生的来源。其中一个来源是样本选择偏差，这可能会导致片面的推论结果；该偏差指的是用一种带有系统偏差的方式从整体样本中选择观测对象。一个明显的例子是只选择那些支持我们理论的案例。样本选择偏差也有其

他更不易察觉的方式。比如,产生偏差的另外一个原因是遗漏变量问题,即没有考虑一些控制变量,从而影响到了解释变量和被解释对象间的因果关系。本书第 2 至第 6 章将依次讨论如何产生无偏推论以及威胁无偏性的潜在隐患。

第二个准则则建立在"效率"这个统计概念基础上:对数据的有效利用涉及最大限度地使用描述性及因果性信息。获得效率最大化不仅需要研究者动用全部数据,更需要获得这些数据的相关信息以提高推论质量。比如,数据是在小地域单位上收集的,研究者就应在与之相应的层次上使用它们,而不需要把它们汇总成国家层面的数据。当然,较小层级的汇总数据意味着更高的不确定性。如果它们是理论的可观测现象,这些数据就同样包含和推论问题相关的有用信息。

1.3 本章的主题

本章概述性地强调了研究设计过程中的四个重要主题,这些问题将在本书的余下部分中得到详细的阐述。

1.3.1 利用可观察的现象将理论和数据联系起来

在本章中我们强调,任何有价值的理论,只有其预测的现象符合观察值时才可能是正确的。我们必须依赖可观测的现象来指导数据收集工作,并且把相关信息从无关事实中筛选出来。本书在 2.6 节中将讨论理论如何影响数据收集工作以及数据如何推进理论的发展。这里需要强调的是,理论和实证研究必须紧密结合起来。任何有用的理论都要对实证调查有指导性意义;不依靠理论来选择问题,任何实证研究都不会成功。理论和数据收集都是检验理论必不可少的过程。实际上,以上两个过程中的推论工作均以不确定性作为特征。

研究者面对任何一个理论时都应该发问:什么是其可观测到的现

象？对任何一个实证调查也该询问观察值是否与理论相关，如果是的话，它们怎样帮助我们判断理论的真实性？在任何社会科学研究中，理论的可观测现象和现实都需要互相配合：只有建立在理论和密切相关的数据基础上，并且经过理论的可观测现象的明确阐述和仔细检查，社会科学的结论才是可靠的。

1.3.2 最大化解释力

试图寻找假设额外含义的学者其实是在探求社会科学中的一个重要目标：用尽可能少的证据去解释尽可能多的东西。高水平的社会科学研究总是设法扩大与解释信息相关的影响。如果我们能用一个或几个变量去准确解释那些重要且复杂的因果效应，对该问题的解释力自然就很高。如果我们能在一个或几个变量的基础上解释众多的效果，解释力同样很高。总的来说，社会科学中的解释效力一般都不会太高，在某些领域中，解释力可能会更低。也许是因为学者还不清楚如何提高它，也许是因为研究对象并不是我们习以为常的事物，也可能两个原因都存在。在传统的定性研究领域中，解释变量的利用率通常都很低。似乎任何解释都需要大量的解释变量：很多证据被用来解释很少的东西。在这种情况下，研究者的目标就应该是设计出更具解释效力的研究。

对于某个特定的研究问题，可能有许多提高解释变量利用率的方法。其中最主要的方法是增加假设的可观测现象，然后对这些现象进行求证。正如上文所述，这项工作包括：(1)改善理论以获得更多可观测的现象；(2)提高数据质量，以观察到更多的现象并以此检验理论；(3)提高对数据的利用率，以便从已有数据中获得更多的现象。与本书1.2.2小节中提到的简约概念不同，这三条和最大化解释效力概念一样，都是研究设计中的规则而不是对真实世界的假设。

最大限度地提高解释效力是非常重要的，当然也非常普遍。因此，我们强烈建议研究者要习惯于列出数据中关于假说的所有可能的现象。在已有数据中也可以检验新的现象——只要这些现象不是来源于

现有数据，而是由理论或其他数据中独立提出的假设。当然，最好的方法依然是求助于其他数据。因此，还应当考虑用在其他数据中出现的现象来检验已有假说，比如来自其他分析单位的数据、关于已有分析单位其他方面的数据、不同层级的数据及不同时间段的数据（比如对将来所作的预测）。发现更多的证据将使解释更具力度，也使得自己和其他学者对获得的结论更加有信心。

对于从最初理论之外的其他来源或汇总数据中去收集观测结果的想法，许多研究者的第一个反应就是反对。例如利伯森（Liberson，1985）采用定性方法对统计概念“生态学谬误”（ecological fallacy）进行研究以提醒人们提防跨级别推论的危险。①该谬误指的是错误地用汇总数据做出个体层面上的推论。使用汇总数据经常会得到关于个体的错误推论：如果我们对个体感兴趣并且数据可得的话，直接研究个体是更好的方法。如果研究者试图获得的推论的解释力不是特别有限，那么理论的有用性就可以体现在很多方面。这个时候，研究者也可以使用多个层级的数据为理论提供信息。因此，即使主要关心的是总体层次上的分析，研究者依然能够通过观察其他级别的数据使得理论拥有更大的解释力。

举一个例子，如果研究者试图解释革命，一方面应该从整体结果中寻找那些可观测的现象，另一方面也可寻求一些具体现象，比如针对某位革命家的深度访谈、一个国家某个地区中一个小社区居民的反应及政党领导人的正式声明等。我们应该尽可能多地从这些地方获取信息以提升理论的可信度。最好的情况当然是研究者能用革命的结果去检验理论。不过多数情况下这一级别的信息量非常少，也许只有一个或几个观察值，同时其潜在价值也不明确或存在测量偏误。此外，会有很多不同的理论可以解释革命这一事实。所以，只有更深入钻研案例或引入其他情况下才存在的相关信息，才能将那些看上去类似的理论区

① 这种表述也许会让读者感到困惑，因为从总体到个体的推论过程既不是生态的，也谈不上谬误。“生态学”仅仅是一个不幸被选中的词汇用来描述基于汇总水平的分析。虽然鲁宾逊（Robinson，1990）在关于该问题的文章中认为基于总体的分析作出关于个体的推论是荒谬的，但很多使用定量方法从事研究的社会科学家及统计学家都认为这种做法有其可取之处。同时，很多可以获得无偏估计的“生态学”方法已经被发展出来。

分开来。

利用其他级别或来源的信息去研究一个关于总体层次的理论存在的唯一问题是,新观察到的现象是否包含与理论相关的信息。如果新增加的观察值有助于检验理论,那么即使它们并不是最有价值的现象,也应该被妥善使用。比如,某些情况下研究者并不关心那些革命家的意见,不过如果他们对问题的回应符合革命理论,那么理论将更有可能正确。由此收集更多的信息也就值得尝试。事实上,对汇总数据的分析观察,比如一个被预测到的革命,仅仅是理论的一个可观测现象。由于其包含的信息量较少,所以它并非一定优于其他可观测的现象。因此,研究者还是应该尽可能多地收集与理论相关的可观测现象。

1.3.3 汇报推论的不确定性

不管在定性研究还是在定量研究中,所有知识及相关的推论都带有一定的不确定性。两种研究风格中的测量都容易发生错误,只是错误的来源可能不同。定性研究中,访问员对一个已经了解其背景的受访者进行长时间的深入访谈,这使得对受访者政治意识形态的测量出错的概率远小于那些对一个由随机抽取选定,且毫不了解其背景的受访者进行结构性访谈的调查(如果过于依赖一个不太可信的受访者,结果甚至可能完全颠倒)。相对那些做深入访谈的研究者,这类研究的研究者也许并不倾向于将个别个案不恰当地推论到整体。不过以上两种方法都会受到测量方面不确定性或某些概率性事件的影响。

所有优秀的社会科学家,无论从事定性研究还是定量研究,都应当汇报对推论不确定性的估计。在定性的政治科学研究中最严重的一个问题是不能合理估计推论的不确定性(King, 1990)。依照本书讨论的规则,即使证据十分有限,研究者也是能够获得有效推论的。只是我们应该尽量避免从那些不充分的数据中得出全面的结论。问题的重点并不是定性研究不可能获得可靠的推论,而是应该始终汇报对每个推论中确定性程度的合理估计。诺伊施塔特和梅(Neustadt & May, 1986: 274)致力的研究很难做出精确的定量估计,于是他们就用一个有效的

方法询问决策者(就是那些需要在没有充分数据情况下也要做政策决定的人)对其结论不确定性的判断。他们问“对于你的结论,你愿意用多少钱打赌”,只要对方反问“赔率是多少”就可以了。

1.3.4 像社会科学家一样思考:怀疑精神以及提出竞争性假说

因果推论中的不确定性使得优秀的社会科学家很难轻易接受它们。当被告知是 A 导致了 B 时,那些“像社会科学家一样思考”的人马上就会质疑 A 与 B 之间的相关关系是否为因果关系。人们倾向于对别人的研究提出这样的问题,而让他人对自己的研究提出这些问题才是真正重要的。即使一个因果解释乍看上去是正确的,我们依然有很多理由来怀疑它。在报纸上看到这样的报道:由于日本人比美国人吃的红肉少,所以日本人心脏病的发病率也较美国人低。这是一个有趣的观察现象。太多的牛排导致美国人中较高的心脏病发病率这个解释看上去也是行得通的。对于那些有怀疑精神的社会科学家,他们就会询问数据的准确性。比如:对两国国民的饮食习惯了解充分了吗?采用什么样本?日本和美国对心脏病的分类是否相同?只有弄清这些,我们才可以对上述现象做出直接比较。即使数据是准确的,依然可以找到其他原因解释这一差别:是否有其他变量(比如饮食习惯、两国国民的基因特点及生活方式),也能解释这样的结果?我们是否粗心地弄反了两者的因果方向?尽管因为心脏病而导致人们少吃红肉有点难以想象,但还是存在这样的可能:人们也许会在晚年对汉堡与牛排失去胃口。如果是这样,那些没有得过心脏病的个体将会活得更长且较少吃肉。这些事实同样导致上述关系,并导致研究者轻率地得出食红肉将导致心脏病的结论。

上段中讨论的内容并不是对此类医学研究表示怀疑。我们的目的是鼓励大家利用社会科学的视角审视针对该问题的因果推论——就是说,始终要带有怀疑态度,关注是否有其他被忽略的解释因素。因果推论实际上是一个过程,后续的研究者可以对之前的结论做进一步的改进与检验。只有这样,我们才能够逐渐接近准确的因果推论。

2 描述性推论

不论是定量研究还是定性研究，都同时包含描述性及解释性目标。一部分学者致力于描述世界，而另一部分学者则致力于解释世界。两种类型的研究都很重要：缺乏好的描述，研究者就无法建构出一个有意义的因果解释；如果与因果解释毫无关系，单纯的描述就会显得索然无味。描述通常在先，因为如果我们不清楚这个世界中哪些特征需要去解释，或者对这个世界一无所知的话，解释也就无从谈起了。当然，描述和解释是相互作用的：解释能够引导研究者观察世界的其他方面；同时，描述也会带来全新的因果解释。

不管是描述还是解释，都需要遵从科学的推论规则。本章的重点放在描述和描述性推论上。描述性工作既不是机械的也不是简单的，它需要研究者从近乎无限的事实中做出细致的挑选。科学的描述包含一些基础性特点：第一，描述包含推论。描述性工作的任务之一是从那些已经被观察到的事物中获取事实去推论那些没有被观察到的事物中的信息。第二，通过科学的描述区分观察事物中的系统性部分和非系统性部分。

对于贬低纯粹描述工作的观点我们并不认同。即使研究的最终目标是由果索因的解释工作，描述在其中发挥的作用也是不可或缺的。不能用描述和解释来区分科学研究，真正区分的标准在于它们是否通过有效的程序得到系统性的推论。不管是描述还是解释，定性还是定量，获得推论才是那些高质量社会科学研究的最终目标。系统地收集事实当然是很重要的工作，离开它科学就无从谈起，但它并不是全部内容。扎实的档案研究和历史事实汇总构成了一部很好的描述史，但是

对社会科学来讲却远远不够。

本章将将区分出描述和描述性推论，前者指的是对事实的收集。2.1 节将讨论这两个看似矛盾的学术目标：发现普遍知识（general knowledge）和了解特定事实（particular facts）。在第 2.2 节中，我们将对推论的概念进行详细的讨论。对于一些观点，除了采用文字表述之外，也会用到非常简单的公式化模型以帮助读者获得更深入和规范的理解。随后的 2.3 节将讨论这些模型的基本特征，接着在 2.4、2.5 和 2.6 节中将依次讨论关于数据搜集、汇总历史细节以及描述性推论方面的模型。在最后的 2.7 节中，将提供评价描述性推论的标准。

2.1　普遍知识与特定事实

社会科学家的研究包含许多特定的组成：比如一个投票人、一个政府代理人、一个城市、一个部落、一个团体、一个州、一个省以及一个国家，等等。好的社会科学研究总是要尝试超越这些特定组成去获得普遍的认识。当然，普遍化并没有削弱特殊性的地位。事实上，从特殊到普遍的转变可以同时加强对双方的理解。一方面，人类社会的某些特定实体，准确地说是关于这些特定实体的信息，为普遍化提供了基础；另一方面，我们总是通过普遍性结论来加强对特定案例的理解。如果研究者已经了解了巴西政府外交部长辞职的原因，这将有助于理解巴西政府其他部长辞职的原因与其他国家外交部长辞职的原因，甚至了解为何一个人辞去他/她在政府或者非政府中的职务。所有这些也会增进对人类行为中不同类型的普遍事实与准则的理解。即使近期发生的巴西外长辞职事件是我们想探求的唯一目标或者目标之一，这些理解依然非常重要。比如，通过研究其他部长行为，研究者可以知道巴西政府部长辞职是为了抗议总统的某些行为，而仅仅研究外交部长的个人行为未必能获得该解释。

一些社会科学研究在没有分析特定事件或某个事物之前就试图对

某一类事件或事物发表看法。研究选举行为的学者采用民意调查来研究人们在总体上的选举决定，而不是关于某个人的选举行为。研究国会财政的研究者在跨社区层面上试图解释金钱对于选举结果的影响。除了顺便提及或者正巧支持普遍性结论，大部分这样的研究都不会单独提到宾夕法尼亚州的第七国会选区或者其他选区。这些研究遵从了普沃斯基和图纳(Przeworski & Teune, 1982)的告诫——“不要出现专有名称”。尽管这些研究的目的不是去了解某个特定选区的情况，但它们也要确保由那些具体选区得到的普遍结论是精确无误的。

其他研究则试图告诉读者关于这些特定案例的信息。比如，研究法国大革命或者其他重要事件的学者就试图解释这些事件为什么发生及怎样发生。如果没有专有名称，基于该传统的研究对读者来说会显得索然无味。即使没有详细观察过一个特定选区或特定候选人，政治学家也可以多次轻车熟路地阐述国会选举中存在的内在规律，但是让我们想象一下如果罗伯特·卡罗(Robert Caro, 1983)没有提及选战的主角林登·约翰逊(Lyndon Johnson)和科克·史蒂文森(Coke Stevenson)，他怎么讨论 1948 年得克萨斯州参议员的选战呢?① 像法国大革命或者 1948 年得克萨斯州的民主党参议员初选之类的特定事件有一定的趣味性和重要性，这些案例激发了我们的好奇心，也是接下来一些重要事件的前提(拿破仑战争或者约翰逊成为美国总统)。为了了解接下来的那些事件，研究者需要研究它们。此外，案例中的那些革命、叛乱以及内战也为研究法国大革命的原因提供了极其宝贵的信息。

本章接下来的内容将讨论上面提到的问题。2.11 小节讨论“诠释”(interpretation)——一个作为科学推论的替代概念；对于研究主题的唯一性和复杂性讨论将在 2.1.2 小节中涉及；2.1.3 小节将介绍关于比较案例研究领域中的一些内容。

① 当然我们不能否认卡罗的记者及传记作者身份，正因为如此，他的目标也许和社会科学家稍有不同。但他的工作中涉及的一些问题也是政治学家所关心的。比如：哪些因素会导致竞选的成功和失败？竞选资金在赢得选举中发挥了什么样的作用？哪些因素可以增加捐助行为？针对这些问题的讨论虽然集中在某个特定选区的特定候选人身上，但这些主题和疑惑也具有一般化意义。

2.1.1 “诠释”和推论

在人文科学中，一些历史学家及人类学者试图通过所谓的“诠释”去寻找那些独一无二的特定知识。诠释学者试图获得关于历史细节的准确总结。他们也试图通过通俗易懂的语言来描述事件以使其他人明白这些工作的重要性。正如费内中(Ferejohn)(见 Goldstein & Keohane, 1993:228)说道:“我们都希望社会科学理论能够给出事件的因果解释……给出原因或阐述某些社会行为的含义。我们不光想知道是哪些因素影响了代理人的行动，也希望了解代理人出于何种个人原因采取这样的行动。”格尔茨(Geertz, 1973:17)对此也表示认同:“在着手研究之前，直接忽略那些有趣的人类行为并不可取。”

那些强调采用“诠释”作为研究方法的学者试图采用“同理心”(“verstehen:德语，意思是站在他人的立场来理解行动与互动的意义”[Eckstein, 1975:81])来解释人类行为的意图。对于那些有意识的行为，诠释学派通过研究与之相关的概念及实践来解释其产生的原因。他们用于评估的标准是:“最明显的标准需要具备一致性并且有一定的适用范围:诠释性解释应该为社会实践提供最大的一致性或可理解性的说明;对某些特定的社会实践，对其诠释也应当和其他社会实践或传统相一致。”(Moon, 1975:173)

诠释学派所提出的最重要的操作性建议是:在提出研究问题之前，研究者应该对研究对象所在的文化进行深入的了解。只有经过很深的文化熏陶并清楚了解研究对象时，研究者才能够提出正确的问题和有意义的假设。在芝加哥一间没有种族隔离的食堂中，杜涅尔(Duneier, 1993)对黑人与白人工人阶级的集体生活进行的研究就提供了可参照的范例。在受过当地文化四年熏陶之后，他发现了几个之前从来没有想过的问题，比如，尽管这些工人很反对共和党，但对于很多议题他们的态度却非常保守。

有些学者将“诠释”这种研究方法捧得更高，甚至认为它是进行社会科学研究的一个全新范式，“社会科学不是一门寻求规律的实验性科

学，而是一门寻求意义的诠释性科学”(Geertz，1973:5)。然而，我们认为科学(对于其定义，见本书1.1.2小节)和诠释并不总是背道而驰的。两者都需要研究者对问题进行细致的描述，对社会拥有较为深入的了解，提出一些高质量研究问题，在更普遍的理论基础上提出一些可被证伪的假设并收集证据来验证它们。只不过科学与众不同的贡献在于它提供了一系列寻找答案的方法及程序去回答那些设计恰当的描述性或因果性问题。

对于“推论”方法论的强调并不是去否定那些促使研究问题形成的过程的重要性。相反，和诠释学派观点一致，我们也认为在提出某种假设或为寻求答案而制订系统研究计划之前，深入了解文化背景十分关键。我们想要补充的是只有依靠科学的推论逻辑，从类似参与式观察法这样的方法得出的结论，其可靠性才有可能被评估。当然，对一个本来就错误的问题，即使找到了正确答案也没有用处。基于“同理心”常常成为很多有见地的假设的来源。比如理查德·芬诺(Richard Fenno，1978)通过他所谓的“渗透和刺探”(soaking and poking)对国会进行的近距离观察就是一个很好的例子。这项研究除了增加我们对制度的理解之外，特别有助于构建好的研究问题。帕特南(Putnam，1993:12)在对意大利人聚居区的研究中就指出：“‘渗透和刺探’要求研究者将自己完全融入到制度的每个细微之处，就像那些每天都在该制度下生活的人一样，去体验它的习俗和惯例、它的成功和失败。如此熏陶能使我们的直觉变得更加敏锐，并为理解该制度是如何成为一个有机体，以及如何适应环境等问题提供无数有价值的线索。”任何关于科学的定义，如果它不能为假设的产生提供空间，就如一个根本不想去发现真相的阐释性描述一般愚蠢。

一旦建立了假设，就需要合理的科学推论来证明它们是正确的，当然也包括对其中不确定性程度的估计。此外，信奉诠释主义的社会科学家采用推理方法的标准也必须与其他定性或定量研究者相同。也就是说，即使我们认同优秀的社会科学需要有见地的诠释以产生有效的假设，我们依然坚持认为科学对于准确诠释是必不可少的。如果只通过“同理心”去了解人类行为，我们就永远不能证伪描述性假设，或为它

们提供自己经验之外的证据。从中获得的结论也永远不能超越那些未经检验的假设,而阐释将止步于个人理解,而非科学的研究。

在诠释主义历史上,一个最好也最有名的例子是克利福德·格尔茨(Clifford Geertz)对吉尔伯特·赖尔(Gilbert Ryle)关于眨眼和使眼色之间区别的分析。格尔茨(Geertz, 1973:6)写道:

> 想象如下场景:两个孩子都快速地眨了他们的右眼。其中一个孩子只是无意识地眨眼,而另一个则是给他的同伴使眼色。这两个行为仅就动作来说没有任何区别;假设摄像机对他们进行"现象主义"(phenomenalistic)的单独观察,也没人能够区分出哪个是在使眼色而哪个仅在眨眼而已,也许两个人都在使眼色或都在眨眼。两种现象之间的区别仍然十分巨大且是摄像机无法区分的,就像一些人将无意识的眨眼误认为是在使眼色那样。使眼色是以一种精确的、特定的方式去交流:(1)故意地;(2)特别地针对某一个人;(3)传递某个特别的信息;(4)按照已为社会接受的行为规则;(5)避免被其他在场的人发现。正如赖尔指出的,使眼色的人其实同时干了两件事情:眨眼以及使眼色。而眨眼的人只做了眨眼这一件。当存在一个约定俗成的认同表示可以用眨眼来传递信号时,这个动作就是使眼色。

格尔茨由此提出了一个重要的概念。如果现有交往理论没有给出"眨眼睛"这个概念,对于那些研究社会关系的学者来讲,即使是"人类眼睑肌肉收缩"这种精确的定量定义也毫无意义。在这个例子中,对于提出"眨眼到底是无意眨眼还是故意使眼色"这样的研究问题,耗时数月的"渗透和刺探"以及从详尽文化研究中提出理论都是至关重要的。通过该例子读者可以清楚地看出诠释的重要性:该方法提供了一个观察社会的新视角,从中可以获得值得思考的新概念以及等待验证的新假设。如果没有深入地融入到情境之中,研究者甚至都想不到哪些是值得验证的理论。在上述例子中,如果没有考虑眨眼和使眼色的区别,研究就根本无法启动。如果通过诠释或者其他方法有助于研究者提出新概念或假设,该方法毫无疑问是有效的。诠释及对文化进行深入理解的其他类似方式的价值将会被反复证明。

在理论上做出眨眼和使眼色的区分之后，研究者需要进一步检验“使眼色确实正在发生”的假设。对该假设的检验需要逻辑清晰的科学推论，本书中所叙述的系统方法是判断眼睑收缩含义的最好方法。假设去区分使眼色和眨眼是研究的核心工作，对此并不难设计出一套研究方案。如果某些特定的眼睑收缩动作被认为是带有政治意义的，那么其他相似动作也可以被观察到。诸如社会行为规则这样复杂的信号传递机制一经形成就会被遵守及采用下去。因此，研究者就可以记录当事人的每一次眼睑收缩动作，观察他的同伴是否在相同时刻发现这个信号及是否给出回应。研究者甚至可以设计一个实验去观察某一种文化背景下的人们是否习惯于这种交流方式。理解背景文化，仔细记录观察到的事件并对相似情形做细致深入的了解，这些都有助于提出正确的研究问题，进而使结论更加可靠。当然，只有依靠科学的推论方法才能够检验假设是否正确。

格尔茨对使眼色的诠释也可以用因果假设表达(关于因果假设的确切定义见本书 3.1 节)：使眼色的行为对行为接受者的因果影响，应该等于接收者看到他/她的眼色后的反应减去没有看到且同时不存在其他变化时的反应。如果眨眼是在使眼色，那么其因果影响将是正向的。如果仅仅就是眨眼而已，就不会存在因果影响。如果要估计该影响的大小(并因此查明眨眼的动作是眨眼本身还是使眼色)并作出推论，我们就需要理解本书以后章节中关于推论问题的详细讨论。

如果被解释成使眼色的眨眼实际上仅仅是没有任何含义的肌肉收缩，那些基于自愿性社会互动理论(voluntary social interaction)的关于眨眼的因果推论将不再适用，因为我们无法将它普遍化。这一点读者应该清楚。[①]

设计区分使眼色和眨眼之类的研究当然不会成为政治学研究的主要部分，但其涉及的方法论问题却时常出现在政治科学的诸多领域中。毕竟研究者经常需要解释某个行为的含义。外交政策的决策者们互相

---

① 为了讨论的完整性，我们可以设想一个完全不同的理论：眨眼虽不是在使眼色，但还是可以对其他参与者有一定的影响。比如一方仅仅是眨眼却被另一方误解。如果感兴趣的是是否有意眨眼，研究者就需要寻找关于该理论的其他可观察的现象。

传递信息。哪些信息具有威胁含义？哪些具有协商含义？哪些仅仅是呼吁本国民众的陈述？对相关的文化规范、国际交往的惯例、某一参与者的历史背景以及关于沟通附加功能(ancillary features)的仔细观察都有助于更好的诠释。试着思考以下定量研究中的难题：美国选民在民意测试中都表现出不愿意参加选举的态度。那么低参与率意味着什么呢？意味着选民与政治体制的疏远，还是选民在计算选举成本与收益后发现付出的成本大于收益？对候选人与选战失望？抑或是选举最低年龄的改变导致这样的结果，还是民调的时机不成熟？因此，公民拒绝去投票就如同使眼色或一条外交信息一样可能有很多含义。老练的研究者应该始终致力于提出正确的问题，然后仔细设计研究以找出那些模棱两可行为背后的真实含义。

一部分诠释主义的支持者甚至提出了更为极端的主张，他们认为某些研究目的就在于不需要观测结果的感知和理解。当然这不是大多数人的态度，但该看似有说服力的论断的确值得进一步讨论。与上述极端观点完全相反，早期狂热支持实证主义的学者认为不可观察的概念在科学研究中是站不住脚的，这些论断在以后的实证研究中同样被证明是不可取的。例如，普萨瑟斯(Psathas, 1968:510)就认为：

> 在任何具体和直接的观察中，只关注那些公开部分的行为至少可以被认定是幼稚的。对于试图理解社会现实的社会科学家来说，他们面临的挑战是去理解参与者的行动对他们的意义。

如果社会科学家只关注那些可观测的行为，往往会忽略其他诸多问题。在这一点上，普萨瑟斯的看法是正确的。可是如果我们什么都观察不到，又该怎样去认知呢？打个比方，假如两种自我认知理论在被观测到的表现形式上相同，那么研究者就没有足够的信息去区分出它们。无论研究者多么聪慧，其文化触觉多么敏锐，阐释技巧多么高超，对个人假设的概括有多么完美，付出多大的努力，其结果都是徒劳的。诠释、感知、深度描述、参与式观察、非参与式观察、深入访谈、移情(empathy)、量化和统计分析及其他所有方法和途径都不能有效地区分这两种具有相同观测现象的理论。假如这两种理论有着不同的观测结果，运用本书中提供的方法就可以区分它们。

在实际研究中,民族志学者(当然也包括其他优秀的社会科学家)确实在搜集可观测现象用以区分他们的理论。也许有的学者更加专注对文化的理解,但无一例外地需要依赖各种形式的观察。对文化背景的进一步"解读"都源于前面提到的观察方法。然而,找到合适的观察法并不是一件容易的事情,甚至可能是整个研究过程中最困难的环节。在那些由定性研究主导的传统调查领域,这一点尤为(也必然地)突出。

2.1.2 "唯一性"、复杂性以及简化

一些从事定性研究的学者也许会反对这种的观点:常识是理解具体事件必要且有用的(至少可能是)基础。对于他们致力研究的课题,他们认为是独一无二的。从某种意义上说,该观点是正确的:的确只有一次法国大革命,也只有一个泰国。凡是读过相关传记或生活在20世纪60年代的人都不会怀疑世界上只有一个林登·约翰逊的事实。不过这些人的观点远不止于此。根据他们的观点,每一种解释都只能局限于某个唯一事件或群体:不去研究革命为什么会发生,只研究法国大革命为什么会发生;不研究民主化为什么会滞后,只研究民主化为什么在泰国滞后;不研究候选者为什么胜选;只研究林登·约翰逊为何能在1948年和1964年的大选中获胜。持这种观点的研究者深信如果给出一般化的解释,比如关于革命、民主化以及参议院选举,他们就会失去解释具体事件的能力。

然而,"唯一性"实在是一个容易使人误解的词。确实,法国大革命、泰国以及林登·约翰逊都是独一无二的。所有现象和事件在一定意义上也是唯一的。法国大革命当然是,但1988年宾夕法尼亚州的第七选区的国会议员选举也是,甚至在那年总统选举时,数百万选民中的每个人的投票决定也是独一无二的。从历史的观点看,社会现实中的每一个方面都无限复杂,并且以某些方式与前面发生的自然及社会事件相联系。因此,内在的唯一性是人类状况的一部分:它并不能以此区分哪些情形可以进行科学的普遍化,而哪些情形不适用。正如本书第

1章讨论恐龙灭绝理论时所提到的，即使是唯一的事件也可以通过关注理论的可观测现象加以科学地研究。

唯一性引出的真正问题是复杂性。关键并不在于事件是否具有内在的唯一性，而在于能否从大量事实中抽象出我们希望理解的社会现实中的关键特点。这正是社会科学研究中首要的也是最困难的任务：简化。它很可能使得研究者遭受类似于“过度简化”以及“忽略重要方面”之类的指责。然而，这种简化工作对所有研究者来说都是必不可少的。简化问题已经成为每一项知名学术研究——无论是定量研究还是定性研究，人类学还是经济学，社会科学还是自然与物理科学——的重要组成部分，也可能永远都是。即使由拥有丰富背景知识的优秀文化阐释者做的最全面描述，都需要尽可能地去简化、具体化并系统地整理所观察到的社会现实。的确，世界复杂性程度与最深刻描述复杂程度间的差异远远超过最深刻描述与最抽象量化或形式分析间的差异。无论多么深刻的描述，也无论其中蕴含着多么丰富的解释，都不可能抓住这个世界上所有的“喧哗与骚动”。因此研究者除了简化之外别无选择。有系统的简化过程对获得有用的知识来讲是非常关键的步骤。正如一位经济史学家所说，对独特性的强调如果“被引向忽视一切规律的极端，社会科学就不会产生。历史学家就会像民谣歌手一样变得毫无目的性”(Jones, 1981:160)。

分析家只有在对历史和文化有了深入理解之后才能简化其描述工作。社会科学家可能仅仅借助事件的其中一部分来进行推论。然而，如果学者试图用一种简化的和科学的方法从事研究，具备丰富的、未系统化的历史与文化背景知识通常是避免错误发生的必要条件。如果研究者对法国大革命或者1948年得克萨斯州的选举知之甚少且漠不关心，就不会有人相信他关于革命或参议院选举的概括成果。

总之，我们认为社会科学研究应该尽量做到普遍与具体并重：它既要告诉读者关于一系列事件的特点，也要提供特定地点发生的具体事件的信息。在某一时点上，我们既希望永恒也希望时间就此凝固。强调普遍还是强调具体取决于研究的目的，不过两者完全有可能并存。此外，这两个目标之间并非相互排斥，而是相互支持。事实上，理解某

个具体事件最好的途径可能正是应用科学的推论方法去研究那些类似事件中的系统化模式。

2.1.3 比较案例研究

政治科学家的大部分工作在于系统描述那些具有重要政治意义的事件。人们关注苏联的解体,关注阿拉伯公众对将伊拉克驱逐出科威特的联合国授权战争的反应及美国最近的国会选举结果。在对这些事件与其他事件(不管是当代的还是历史的)的联系上,政治学家的研究往往比新闻报道更加全面,因此在这方面,公众更加信赖政治学家。对事件的描述应该尽可能准确和系统化。如果可以量化那些我们试图了解的事物,那就尝试着应用:有多大比例的苏联报纸批评政府政策?在约旦和埃及实施的民意测验揭示出约旦人和埃及人对海湾战争所持态度有何不同?多大比例的现任国会议员将有希望赢得下一任选举?

量化工作可以获得精确性(precision)但未必能获得准确性(accuracy),因为研究者需要引入许多与被测量概念或事件并不紧密相关的定量化指数,这样做可能导致严重的测量误差及因果推论问题(对此点的讨论见本书第 5.1 节)。此外,对于一些本来就很难量化的时间,研究者也很难找到精确的方法去描述。那些训练有素的定性研究者试图仔细分析宪法和法律,而不能泛泛而谈。在针对政府政策的案例研究中,研究者会问受访者一些尖锐的、有针对性的问题,这些问题的答案往往比较明确。同时受访者的即兴回答也可以启发其他假设。对于描述工作来讲,案例研究必不可少,同时也是基础性的。在尚未给出一定程度的精确描述前就尝试解释往往没有什么意义。

对复杂事件进行有洞察力的描述非常有价值。描述性工作在比较政治学或是国际关系学中显得尤为重要。因为有许多事物有待我们去了解,也由于已有研究的解释能力并不够强,还因为好的描述往往依赖于好的解释。在一定程度上,需要了解的事物与解释中存在的缺陷往往基于相同的原因。譬如在最近的国际政治中,政权的类型、结盟、国际上的相互依赖都在发生急剧的变化,这不仅改变了国家间相互作用

时所处的背景，也进一步增加了描述新形势的需要。国家和各方参与者都试图预见对手的行动并做出反应，这往往导致因果关系难以建立。因此在解释国家行为时，预期的行动和已观察到的行动同样重要。在解释国际政治时，如果忽视了战略互动和预期反应的话，还不如去细致地描述那些重要的且相互关联的事件。毕竟，好的描述也远胜于差的解释。

有一个关于深度案例研究方法的优点常常会被大家忽略：发展出一个好的因果假设对好的描述工作来说是一种补充，而非它的对立面。即使某个研究最终都没做因果推论，研究者从一个需要解释的问题构造出的案例研究中也能获得集中且恰切的描述。

如果遵循本书提供的方法，从比较案例研究中完全可以获得有效的因果推论。尽管在当前实践中它们还达不到有效推论的标准（这点将在本书第 3 章讨论）。被大多历史学或阐释主义导向的社会科学家称为“解释”的工作，由于没有达到因果推论的标准，因此仍然停留在描述阶段。从这个角度来说，许多学者认为比较案例研究必须在描述及解释方面更加系统，这一建议十分重要。

比如亚历山大·乔治（Alexander George）就建议采用“结构的集中比较”（structured, focused comparison）这一强调规范性的方法去收集数据（George & McKeown, 1985；另见 Verba, 1967）。乔治与他的合作者也强调要从那些精心选择过的样本中系统收集相同的信息，也就是那些相同的变量。如果能获得因果推论，研究者还需理论的指导以获得系统性描述。①

这种结构化的集中比较方法也就是乔治和麦基翁（McKeown）称之为“一致过程”（congruence procedure）的系统方法。通过该方法，研究者“在案例研究中对数据要求确立定义和标准……进而通过构建普遍性的理论问题来指导每个案例”（George & McKeown, 1985:41）。乔治和麦基翁的如下观点（1985:43）也被广泛采用：“即使对小样本进

① 关于比较案例研究的文献非常丰富。相关的优秀著作有 Eckstein(1975)、Lijphart(1971)及 Collier(1991)。

行控制比较，也需要遵循系统的数据处理过程。”如此“结构化、集中比较”需要收集各个研究单位中相同变量的数据。因此，它与本书所极力强调的描述性案例研究中应用到的那种系统化信息方法并无二致。利用两种方法都可以进行描述性或因果性的推论。此类针对比较案例研究的建议虽然是最基本的，但非常容易被研究者所忽略。

2.2 推论：数据搜集工作的科学目标

推论是利用已知事实去探索未知事实的过程。未知事实是研究问题、理论和假说。而已知事实（不管是定量还是定性）是我们已经收集到的数据或观察值。

在追求一般知识时，不管我们是纯粹为了该目的还是去更好地理解某个事实，都需要避免被那些潜在的或自相矛盾的观点所迷惑。所幸的是，解决该问题的方法恰好就存在于探索一般知识的过程中。也就是说，组织事实最好的科学方法是将其视为理论或假设的可观察现象。在科学的简化过程中，如何有效选择待检验的理论（或假设）非常重要。接下来，理论将引导我们对其揭示的事实进行筛选。把事实作为某一理论的可观察现象组织起来，对研究设计和操作来说也是非常重要的。首先，按照这种标准筛选事实，研究者可以发现增加观察值将有助于检验理论。既然信息越多越好，我们就不应该舍弃这些数据，研究过程也因此得到改进。

其次，在搜集数据前并不需要研究者有一个完整的理论，当然也不需要自始至终地固守一个理论。理论和数据是相互促进的，它们之间的关系就像鸡和鸡蛋一样：有些理论必须在数据搜集前形成，而有些数据需要在理论创建前就要进行收集。虽然教科书说的是用数据去检验理论，但从数据中汲取信息与检验初步的理论及假设一样都是研究的重要目标。汲取信息的过程涉及将数据重新组织成为新理论的可观察现象。该过程在研究工作初期十分普遍，通常发生在数据初步搜集工

作完成后。为了检验新理论，数据收集在重组后仍需进行下去，这有助于避免在构建理论和检验理论时使用同一组数据。①

第三，强调收集信息作为理论的可观察现象，会使定量研究与定性研究具备共同的基础。事实上，一旦不再狭隘地甚至天真地考虑案例、单位或数据层级，研究者就会发现大部分定性研究都会给理论提供了很多可观察的现象，只是其中很多都被研究者忽略了。把这些资料整理成理论的可观察现象列表有助于揭示定性研究的基本科学目的。从某种意义上讲，这相当于在引导一个正在研究某个问题（也可能是某个政府决策）的学者提出这样的问题："假设我正确地解释了政府如此决策的原因，那么真实世界中还有什么其他发现？"研究者完全可能从其他决策过程中发现额外的现象，也可以从正在研究的决策的其他方面去发现。例如，决策是何时做的？怎样做出的？依据是什么？指导理论构建与资料收集工作最重要的原则是：去寻找更多的理论可观察现象。

当一个新的理论或假设被提出来时，把理论揭示的、原则上也能被观察到的所有现象都罗列出来是一个很有效的方法。罗列的内容可以是那些已经找到或很容易找到的现象，这样就能为研究提供基本的操作指导。如果多搜集一份资料能多提供一种验证理论可能的话，那么（在通常的时间、资金及人力约束下）这就值得我们如此去做。如果一个访谈很有趣却不是一个理论潜在的可观察现象的话，它就对于验证理论没有帮助而应该被放弃。

将资料整理成理论的可观察现象是一个简化的过程，这就需要研究者将资料系统化。将那些来自真实世界的原始材料按照"分析单位"或"案例分析"组成的集合转变为按照"属性"或"变量"组成的集合。如果集合是"选民"的话，分析单位就是由选举区中的选民组成的样本，而

① 例如，库姆斯（Coombs，1964）就指出，有益的数据搜集工作需要或者其本身就是某种理论或"小理论"（mini-theory）。虽然研究者在收集大多数定量数据和定性历史记录时就知道其用途，但它们还有一个明确的目标：鼓励后续研究者使用。大家读 15 分钟的《美国统计摘要》（*Statistical of the United States*）就会明白这点。当然，对上述信条遵守的程度不同，不同研究者为数据搜集所付出的努力也会不同。

属性或变量则可能是收入、党派身份或者任何待验证理论所揭示出的可观察现象。集合也可以是一种特定集体，比如社会或国家，具体单位就是对这些集体的选择。属性和变量则可能是它们的规模、政府类型、经济环境、种族构成及其他研究者关心且可以被测量的因素。在收集数据的时候，这些概念与其他各种各样的建构，比如类型、构架及任何形式的分类或分级都是很有效的暂时性工具。不过研究者无法凭借它们获得待检验的假设。一般来说，我们并不鼓励研究者通过这种方式组织资料。相反，研究者需要的是去组织从理论承袭而来的那些概念，这就意味观察到的现象要么是理论的预期现象，要么与理论毫不相关。如果是不相关或不可观察，这些现象就应被忽略。反之，就应该妥善利用。此外，也不是所有的数据都来自同一个分析层级。来自不同时期或不同地域的分散数据或者观察报告都可能为理论提供额外的可观察现象。也许研究者对这些辅助性的现象不感兴趣，但如果它们与理论的预测一致，也是有助于该理论的有效性及适用性的。数据也不都是“对称的”(symmetric)：研究者可以对一个省进行详尽地研究，对两个国家进行比较研究或个别访谈一个政策部门的领导，甚至也可以采用定量方法，只要这些数据是理论的观察结果就可以。在这个过程中，研究者将从“特殊”走向“一般”，因为在共有的基础上对特殊研究单位的描绘就是一个普遍化过程。基于此，我们对普遍知识和特定事实的了解就变得更加深入。

一般来说，研究者都希望获得尽可能多的信息去支持假说，这意味着需要很多额外的案例。不过这往往是很难、耗时或者费钱的。出于这种考虑，我们就应该避免引入那些不相关的信息。例如，逐月记录英国众议院里保守党所占席位，而不是按照历次大选结果记录，将大大增加观察值的数量。不过没有什么用处，因为该做法获得的新信息很少。相反，将美国总统大选的结果具体到州，或者到县的层次，在大大增加案例数量的同时也会加大关于该问题的信息量。

如果研究目的是去了解某个竞选者在总统竞选中的成功原因，那么该研究在本质上属于汇总层级上的问题，研究者就不能采用地方政府层级的信息。不过有些影响总统选举结果的因素确实存在于地方政

府层面上。举个例子来说，如果根据失业率这样的经济变量来预测选举结果，和使用全国性的总体失业率数据相比，使用州一级失业率能为理论检验工作提供更多的观察值。通过证明理论在其他情形下同样成立，哪怕这些情形并非我们的直接关注对象，也会使研究者更加确信理论的正确性。也就是说，研究者准确地解释了一个理论的可观察现象。

2.3 关于定性研究的公式化模型

模型是对世界某些方面的简化和近似。尽管好的模型可以提取那些“正确”的特征来描绘现实，我们却并不能以“真的”或“假的”来评价模型。

想象一个 6 英尺的塑料飞机模型。这个模型只是真飞机的微缩——没有可移动的部分，不能飞行也不能容纳乘客。没有人会把模型与真飞机混淆起来；询问这个模型的真假就像去询问充当蒙娜丽莎(达·芬奇的画作)的模特是否真有如此动人的微笑。即使真有，我们也不能指望达·芬奇的画作是对她本人的精确描绘，不管是描绘模特还是描绘圣母玛利亚。因此，我们自然也不应指望飞机模型能够反映真飞机的所有特征。应当关注的是这个模型是否针对某个特定的问题，是否抽取出真飞机的正确特征。如果意在给孩子一个关于飞机的概念，这样的模型就已足够。如果这个模型是全比例缩小，那么它就有助于飞机设计师进行风洞测验。这个塑料模型抽取的是飞机的形状。出于某些目的，这个特征正为我们所需。该模型当然忽略了飞机的许多细节，比如尺寸、颜色、乘坐感受、各个部分的力量、客舱里的座位数、引擎的动力、坐垫的质地，以及电子设备、空调、管道和其他很多关键系统。当然，如果研究者想了解这些部分，只要建立起相应的模型就可以了。

如果不清楚研究对象的哪些特征是值得关注的，就无从评价该模型了。举个例子，我们也许并不清楚一个旨在研究飞机上灰尘数

量的模型有什么用处。对于教育孩子和风洞测试来说,这样的模型也许根本没用。不过,如果地毯上的灰尘使飞机变重而消耗更多的燃料的话,该类模型对航天工业来讲就至关重要,一旦建成就能节省数百万美元。

所有模型都可以分为限制性和非限制性两种。限制性模型更为清晰、简约、抽象,不过不够符合现实(除非模拟的现实也很简约)。非限制性模型更为详尽,更加注重情境当然也更符合现实,不过却不够清晰,也难以精确估算(见 King, 1989:2.5)。在这个背景下,研究者就需要根据应用目的以及研究问题的复杂程度去建立相应的模型。

有些模型是物理的,而有些则是图示的、言辞的或者代数的。比如,一本定性描述欧洲司法系统的著作也是关于该问题的模型。无论描述得多么细致,作者多么才华横溢,与现实的司法系统相比,书中描绘的内容永远是抽象和简化的。由于抽象化是理解的必要条件,所以是不是一本好书的标准既在于它包含了什么,也在于它剔除了什么。

定性研究者经常使用语言模型,而以下的内容将会讨论如何用公式化模型去改进语言模型。与关于玩具飞机和法国大革命的模型研究一样,关于定性研究的公式化模型不应该与定性研究本身相混淆。这些模型的目的是避免某些问题并提供可供探索的机会。此外,模型还能帮助研究者发现一些意想不到的新思路。

尽管接触过统计模型的读者理解以下模型没有任何困难,但我们仍然假定读者没有公式化模型的训练。这些模型的推论逻辑既适用于定量研究,也适用于定性研究。定量研究者可能更加熟悉这些术语,但这不意味着他们就比定性研究者更善于运用科学的推理逻辑。进一步讲,与定性研究相比,这些模型也并非更适用于定量研究:它们对这两种研究风格来说都是非常有用的抽象化过程。为了使这些代数模型简单易懂,我们也会采用文字描述和代数标记相结合的方法。尽管我们不鼓励读者跳过这些代数符号,不过即使没有这些内容,也不会影响本书内容的连贯性。

2.4 关于数据搜集的公式化模型

在对描述推论与因果推论这两个社会科学研究的主要目标进行讲解前,我们需要为数据搜集与汇总建立一个模型。这个模型非常简单,但可以有效地处理推论中的问题。当然,这里使用的公式化模型没有统计学中的那么正规,不过这并不影响对理论的清晰表达。数据搜集有多种方法,包括观察法(observation)、参与观察法(participant observation)、深入访谈法(intensive interviews)、大规模抽样调查法(large-scale sample surveys)、二手资料调查法(history recorded from secondary sources)、随机实验法(randomized experiments)、民族志研究法(ethnography)及内容分析法(content analyses),等等。数据搜集最重要的原则是要弄清楚数据的来源及获取数据的过程。通过该过程收集到的每一个信息都应当有助于研究者厘清理论的可观察现象。如果这些信息不是目前研究课题的可观察现象,即使它有助于发现新的研究问题,也无益于现有问题的解决。

本书使用"变量"、"单位"及"观察值"这几个概念来建立模型。举个简单的例子,我们打算列出四个人各自的年收入。用数据表达可能只是四个简单的数字:9 000 美元、22 000 美元、21 000 美元和 54 292 美元。通常情况下,我们会将四个人(编号分别为 1、2、3 和 4)的收入标记为 y_1、y_2、y_3 和 y_4。就像度量两个访谈对象各自配合程度可以被赋予"参与"、"合作"或"不合作"等值一样,而且也可能会被标记为 y_1 和 y_2。在这些例子中,变量是 y,单位是个人,而观察值是每个单位(以美元计算的收入或者配合程度)的变量值。符号 y 之所以被称为变量,是因为它的值会随着单位的改变而改变。通常,变量是指那些可以代表随着特定单位变化而改变其值的任何事物。鉴于我们可以超越时间或跨越空间去收集信息,所以单位设计的范畴就十分广泛:个人、国家、组织、年份、选举事件甚至十年的时间都能被作为单位。这些单位有时也可以相互组合形成新的单位。而观察值可以是数字、文本、图

像及任何其他类型的实证数据。

举个例子，假设我们有兴趣去了解 1945 年以后国际组织的变化情况。在收集数据前，首先要弄清楚我们想解释什么。这有很多议题可供选择：可以研究 1990 年国际组织活动的规模分布情况（通过活动范围或者组织本身来了解），也可以研究 1945 年以来国际组织活动规模的变化或 1945 年以来国际组织活动分布规模的变化，等等。关于国际组织活动的变量可以包括在某一时点其成员国的数量、该组织开展活动的次数或其预算及工作人员的规模。在上述例子中，分析单位可以是国际组织、事务范围、成员国情况，也可以是年数，如五年、十年这样的时间段。在数据收集阶段，应该收集哪些变量？应该收集多少？单位数量是否一定要超过变量的数目？变量测量的精确程度，等等。诸如此类的问题并没有明确指导原则可以依靠。唯一可凭借的就是研究者对各个问题重要程度的判断。当对数据用途有一个清晰的思路时，指导原则就是尽可能多地寻找理论的可观察现象。正如本书第 1 章中强调的，实证研究既可以用来检验假说，也可以去启发新假说。如果研究者的目的是后者，则必须用新数据去检验。

读者可以从以上讨论发现，案例研究一般包含很多由不同单位测量而来的变量。尽管案例研究用到的案例一般不多，但其包含的观察值数量却很多。因此，区分出案例的数量和观察值的数量就显得十分必要：前者出于对研究目的的兴趣，但只有后者才能用来判断研究中用以检验假设的信息量，因而后者更加重要。我们通常使用 n 来表示观察值的数量而非案例的数量。只有在非常特殊的情况下，比如各个观察值之间存在相关关系时，我们才会去区分信息量与观察值的数量。观察值数量的表示法源自调查过程中的抽样：在抽样时用 n 表示受访人数。现在只是把该用法的适用范围一般化。本书对“观察值”的定义与哈里·埃克斯坦（Harry Eckstein，1975：85）对“案例”的定义是相似的。埃克斯坦提出：“六次针对英国下议院选举的研究可能是 $n=1$ 的研究，也可能是 $n=6$ 的研究，还可能是 $n=120\ 000\ 000$ 的研究。这取决于研究对象是选举制度、具体选举活动还是选民。”“只有超越了这些个体，关注这些个体的度量，才能消除关于‘个体’（进而‘案例’）是由什

么构成的困惑。”埃克斯坦的文章问世以后，学者们一直沿用“案例”来代表整个案例研究。虽然这个词本身缺乏精确的定义，但本书依然按照惯常用法在可能的地方都使用“案例”。当然有些地方也保留了“观察值”这个概念：该词是指对一个单位的一个或多个变量的测量。

本章剩余的部分将讨论在研究设计中变量及单位是如何提高思维的清晰度的，当然有时候用定量方法概括信息并不总是合适。以下部分试图回答如下问题：如何在做“如历史本来面目”的描述性推论时不至于被淹没在无关细节的海洋中。换句话说，如何从细枝末节中抽取要点。

2.5 概括总结历史细节

在数据收集完后，研究者需要对数据进行概括。概括可以描述那些包括大量数据的资料，但它与推论并没有直接关系。对有待解释的事实进行概括是研究的起点，但鉴于我们对普遍化解释更感兴趣，所以概括并不是最终的目标。

在研究中，概括是必不可少的。当然研究者也不能对所要了解的事件描述起来没完，这样做也没有什么意义。优秀的历史学家都应该知道哪些是关键信息，所以著书立说时强调重点而不会偏离主题。要了解 19 世纪前 15 年的欧洲历史，研究者就需要像拿破仑一样了解军事战略原则，甚至还要知道他的军队在行军途中的伙食。而诸如拿破仑的头发颜色、他喜欢吃煎鸡蛋还是煮鸡蛋，这类问题就无关紧要了。好的历史学著作应该包括对大量历史细节言简意赅的总结，当然仅仅有这些仍然不够。

概括历史细节过程的模型被称为“统计描述”。统计描述是对数据的缩略表现，其目的是以更简单的格式显现出数据的主要特征。①例

① 对包含 n 个观察值的变量 $y(y_1, \cdots, y_n)$ 进行测量，h 统计量就是一个实值函数(real-valued function)，其定义为：$h = h(y) = h(y_1, \cdots, y_n)$。

如，样本均值或平均值的统计描述为：

$$\bar{y}=\frac{1}{n}(y_1+y_2+\cdots+y_n)=\frac{1}{n}\sum_{i=1}^{n}y_i$$

其中，$\sum_{i=1}^{n}y_i$ 是 $y_1+y_2+y_3+\cdots+y_n$ 的一种简便表述法。另一个统计数据是样本最大值，记作 $y_{\max}$：

$$y_{\max}=Maximum(y_1,\ y_2,\ \cdots,\ y_n) \tag{2.1}$$

2.4 节中的例子给出了四个收入：9 000 美元、22 000 美元、21 000 美元、54 292 美元。可以计算出它们的样本均值是 26 573 美元，样本最大值是 54 292 美元。研究者可以利用这两个指标来描述原始数据。当然，研究者也可以据此计算其他样本特征，比如最小值(minimum)、中位数(median)、众数(mode)或方差(variance)，等等。

上述每种概括都试图将所有数据(这个简单例子中是四个数字，另一个例子中是对于欧洲历史某个时段的认识)简化为单一数字。对读者来说，这些统计描述所传达的信息要比原始数据更加简明和有意义。当然，如果一个数据集仅含有四个数字，用五个不同的统计数据去概括它们的特征就没有必要了，原始的四个数字反而更加简单。一般来讲，解释统计数据通常比解释整个数据简单，但是将一组庞大数据概括成几个数字难免会丢失一些信息。

对历史细节的概括需要遵循哪些原则呢？第一条原则是应该将概括的重点放在想要描述或解释的结果上。如果想了解一般国际组织的发展状况，那么将概括的重点放在联合国上就是不明智的；不过如果对不同大小的国际组织的规模分布感兴趣的话，联合国当然是应当集中关注的对象之一。联合国不是一个有代表性的国际组织，但毫无疑问它是一个重要的国际组织。用统计术语讲，如果要调查那些典型的国际组织，就应考察(预算、任务、成员数等)平均值；如果要了解其活动范围，就应考察其方差。第二条原则是概括必须简化那些供我们使用的信息。从定量角度来说，这意味着概括性统计数据的数量一定要小于原始数据中的样本个数。否则，直接依靠原始数据反而可以更简便地

传达信息。[①]也就是说，为了帮助读者理解，概述应当足够简明。任何现象都不可能被完美地概括，所以判断概括是否充分取决于研究目的及读者。比如一篇关于战争与联盟的论文中的数据可能有 10 000 个观察值。作者可以用 50 个统计指标概括整个数据集，即使这样，多指标对于一名专家来说也十分很费解。所以，要是在大学本科的课堂上，讲授者最好用不超过三个图表来描述整体数据特征。

2.6 描述性推论

描述性推论是指在一系列观察值的基础上对未观察的现象进行理解的过程。举例来说，假设研究者有兴趣了解英国保守党、工党以及社会民主党在 1979 年地方选举中的力量消长。基于此，研究者实际观察了那年 650 个地区的众议院选举。

如果我们认为通过记录保守党在不同区域的选举票数以及在整体中所占的席位比率，就能直接观察到该党的选举实力，这种想法是幼稚的。与社会生活和科学探索的所有其他方面类似，政治领域中的随机及不可预见性同样不可避免。[②]假设 1979 年英国议会突然反常地(或许是出于尊重社会科学的考虑)同意每周选举一次，同时假设每次选举互不相关(事实当然不可能这样)。在这种情况下，即使保守党的潜在支持人数保持不变，每周都举行的重复试验也并不能保证每一地区中的每一党派能获得相同票数。这可能是因为天气变化、流行病突然爆发，也有可能是因为选民休假。所有这些事件都会影响选民的投票数量与选举结果。另外，国际上发生的某些偶然事件，比如丑闻的曝光，虽然可能没有长期影响，却同样影响每周的选举结果。所以，众多暂时性事件都会对选票数量产生轻微的影响。因此，通过观察任何一次选

① 这点与本书 4.1 节中探讨的不明确的研究设计紧密相关。

② 参见波珀(Popper，1982)对不明确性的详尽讨论及辩护。

举都不能完全衡量保守党的真实实力。

再举另外一个例子，假如我们试图研究以色列占领的约旦河西岸区域内以色列人（警察和居民）和巴勒斯坦人之间的冲突。由于双边的官方报道都经过审查或删减，因而关于冲突程度的报道都是不可信的。所以研究者决定亲自调查：通过深入访谈或参与家庭与集体活动来确定冲突在不同社区中的水平。如果研究者在各个社区进行为期一周的调查，那么社区的平均冲突水平将部分是该周碰巧观察到的冲突水平的函数。即便将调查时间延长至一年，哪怕冲突水平的不确定性会因此下降，仍然不可能清楚地获得冲突的真实水平。

从以上例子中可以看出，无论是保守党在不同区域中选票的差异还是西岸社区中的冲突水平差异都可以分为以下两个部分：系统性差异及非系统性差异。选民中的系统性差异包括各选取基本的和可以预测的因素，比如意识形态、收入、竞选活动组织的差异及各选区党派偏好的历史传统。在假设进行的每周选举试验中，系统性差异是始终存在的，不过像天气变化导致选票数量变化之类的非系统性差异却是变化的。而在约旦河西岸社区冲突的例子中，系统性差异主要表现为犹太人和巴基斯坦人深层次的文化差异、对彼此的认识、地理上种族隔离的居住模式等。即使研究者多次进行这样的周调查，社区间的系统性差异始终会影响被观测到的冲突水平。不可预测的非系统性差异包括恐怖事件、以色列警察的暴力事件等却是不可预设的，它们对冲突水平的影响仅限于事件发生的那一周。利用适当的推论方法，即使真实数据中存在非系统性或随机差异，研究者还是能够找到本质上的系统性差异的。

因此，进行推论的基本目标之一就是要区分研究现象中的系统和非系统部分。系统部分并不比非系统部分更重要，因此我们不能仅关注其中某一个而忽略另外一个。然而，将两者区分开来又是社会科学最基本的任务。理解推论的一种策略是把所搜集的数据看作众多可能数据中的一个。例如，英国 1979 年的选举结果只是众多可能发生的结果之一，对社区某一周的观测结果也是多次调查的结果之一。

在描述性推论中，研究者需要了解其获得的调查资料在多大程度

上是一个典型现象还是一个极端现象。如果 1979 年英国选举期间爆发的一场流感仅席卷了工薪阶层,却没有蔓延到那些高收入人群中,那么由此得到的观察值就不能度量潜在的保守党实力。因为数据中的非系统、偶然性的要素超过或扭曲了系统性要素。如果在以色列入侵南黎巴嫩之后研究者马上着手观察,其得到的结果同样不能反映约旦河西岸社区中的冲突程度。

理论上,每一个研究问题都能从真实世界中获得多个数据集合,即使这样也不总能满足政治科学家的需求。但在通常情况下,能观察到一组数据就已经很幸运了。根据建模的需要,我们用变量 y 来代表这组数据(比如,工党的选票数),y 会根据地区数 $n=650$ 分别取值(n 表示不同选区):y_1, y_2, …, y_n(y_1 表示在第一选区中工党获得 23 562 张选票)。这组由 y 所标记的"观察值"是实现了的变量(realized variable),其取值在 n 个不同的单位中都不同。我们将 Y 定义为随机变量(random variable),这表示在重复性假设下,其取值会根据选举的不同结果而变化。也就是说,y_5 表示第五选区中投票选举工党的人数,而 Y_5 则是一个随机变量,表示在情况不变的条件下在第五选区举行一次重复选举试验中工党所获的选票数。在所观察的这个样本中,由于各种系统因素、随机因素在各选区中有所不同,工党所获的选票数 y_1, y_2, …, y_n 也随之变化。由此,为了区分这两种形式的"变量",我们通常用 y 来表示"实现了的变量",用 Y 来表示"随机变量"。

同样的建模方法也可以应用于定性研究中。当然,我们不可能也不去打算量化以色列人与巴勒斯坦人之间关系的紧张程度。部分原因在于"冲突"是个复杂的概念,牵涉大量个体感情、团体反对、意识形态冲突等因素。在这种情况下,y_5 是一个实现了的变量,代表在比雷赫(El-Bireh,耶路撒冷以北 14 千米的一个城市)地区,研究者在一周内观测到的全部冲突数量。①随机变量 Y_5 既代表在比雷赫所观察到的,也代表在比雷赫应该可以观察到的偶然事件,该事件的随机性来自进行

① 当然这同样适用于所有其他社区。

观察的周中发生偶然事件的差异。①

推论的目标是研究随机变量 Y_1，…，Y_n 的系统性特征。此处读者需注意"随机变量"与"系统性特征"这对看似矛盾的术语。一般来说，研究者希望区分数据中的系统与非系统部分。当然有时候研究者也想剔除出系统部分而获得非系统部分。比如，我们可能希望了解第五选区中工党选票数的期望值（在该地区进行大量重复选举中工党的平均选票数 Y_5）。该数值表现了现行选举制度的系统特征，所以社会科学家对此很有兴趣。相比之下，观测到的某一次选举中的工党选票数 y_5 则不具有长期吸引力：它是系统性特征与随机误差的函数。②

冲突水平在西岸第五社区比雷赫中期望值（系统要素的特征之一）的公式化表达：

$$E(Y_5)=\mu_5$$

其中，$E(\cdot)$表示期望表达式。从该公式可以得出在第五社区进行每周一次无限重复试验中获得的观测结果的平均值。参数 μ_5（希腊字母 μ 下面加上下标 5）则代表该区中巴以冲突程度期望值的计算结果。该参数是关于随机变量 Y_5 系统性特征模型的一部分。可能有人将观察到的冲突程度 y_5 作为 μ_5 的近似值。但需要读者注意的是，除了该系统性特征之外，y_5 还包含了大量随机因素，因此通常存在更好的估计结果（见本书 2.7 节）。

我们希望得到的另外一个系统性特征是约旦河西岸社区中的平均冲突程度，其计算公式如下：

$$\frac{1}{n}\sum_{i=1}^{n}E(Y_i)=\frac{1}{n}\sum_{i=1}^{n}\mu_i=\mu \tag{2.2}$$

μ 是所有样本中观察到的冲突程度平均值 $\bar{y}$ 的估计值，但关于此系统

---

① 读者需要注意，随机性并不完全和不同观察周间的差异相同，因为偶然事件和系统性差异都可能导致观察结果出现偏差。因此研究者需要假设一个理想化的情形：世界是由恒定的系统性因素及变化的偶然因素组成的。

② 当然，y_5 可能对那年该选区的居民十分重要，因而不管是随机还是系统成分，都值得我们研究。尽管如此，研究者仍应尽力区分出随机部分和系统部分。

性特征也存在其他估计值。(需要注意的是,本书在 2.5 节讨论概括历史数据时,该指标被用来做描述性推论)。另外,关于随机变量的其他系统特征(包括方差等)将在本书 3.1 节中介绍。

除了上述指标之外,我们对同一社区内的系统性特征在数周时间内的变异程度也感兴趣。也就是说,非系统性部分的规模同样值得关注。该指标是通过每个社区的方差(以此替代期望值)计算的:

$$V(Y_i)=\sigma_i^2 \tag{2.3}$$

在这里,σ^2(希腊字母 sigma)表示对随机变量 Y_i 计算方差所得的结果。住在约旦河西岸一个以色列和巴勒斯坦矛盾高度集中的社区里当然很不幸,但住在一个异质性很高同时无法预期冲突程度的社区里也许会更糟。当然,无论哪一类社区都可能引起学者的兴趣。

为了更好地理解这个问题,研究者就需要区分随机变异的两个基本观点。①这两个观点实质上位于一系列观点组合的两个极端。尽管分别有相当数量的支持者,但大部分政治学家的观点还是介于其间的。

> 观点 1:一个完全或然的世界。随机变化普遍存在于自然界、人类社会及政治领域中并且永远不可能被消除。即便可以毫无误差地测量出所有变量,获得一份普查数据(不仅仅是一个样本)且含有所有想到的解释变量,研究者仍然不可能得出完美的预测。因此,将这个世界划分为系统和非系统的两个部分可以在一定程度上改善预测。需要指出的是,研究者在数据分析时所作的任何努力都不能减少现实世界中普遍存在的非系统部分。
>
> 观点 2:一个确定性的世界。随机变异只是世界中那些无法解释的部分。系统性变化和随机变化的区别可以由研究者区分,同时这也取决于解释变量是否可得且被分析采用。只要有合适的解释变量,世界完全可以被预测。

① 对这两个基本观点的详尽论述见金(King, 1991b)。

在不同研究领域中，上述两个观点确实给推论带来某些困惑。①即便如此，在多数研究中，从观察本身来说是等价的。尤其根据观点 2，假设至少有一部分解释变量是未知的，当这些未知变量成为观点 1 中随机变异的解释时，观察上的等价便会出现。由于缺乏任何可观察的现象来区分这两个观点，其间的选择往往取决于个人信仰而非实证检验。

再举一个例子，在两个观点中，判别一个特定政治或社会事件是一个系统性过程还是非系统性过程，将取决于研究者自身。如果采用观点 1，我们可以区分出一个效应是系统性的还是非系统性的。但除非有另一套数据(或仅仅是另一个案例)能证明该效应确实存在，否则做出正确的判断是很难的。

如果采用观点 2，研究者能做的只有描述数据了，因为“不正确地”将事件判断为随机还是系统性的是做不到也是没有用处的。当然关于该视角的一个更现实的版本则是接受观点 1 的判断：某个时间是随机的还是系统的。但这种调整就需要研究者自主地决定哪些问题需要检验，而哪些问题则无法解释。这就意味着在开展分析前，研究者都需要假定所有观测结果均由非系统因素所解释。下一步的工作则是提供证据证明特定事件或过程是系统性作用的结果。对于那些无法解释的事件或过程，究竟是随机导致还是由未知解释变量所导致就成为将来的研究课题了。

上述观点对定性研究和定量研究同等适用。定性研究通常是历史性的，但只有当它作为一门可以明确进行推论的社会科学时才最有用处。要将随机变量概念化并尝试估计它们的系统性作用而不仅仅是概括历史细节的话，并不需要大量数据。事实上，一个优秀的历史学家能

① 经济学家更倾向于第一个观点，而统计学家则倾向于第二个观点。同时，第一个观点也被致力于“质量控制”的工程师们所广泛接受。关于这两个观点的差异甚至在量子力学领域中也引发了激烈的争论。第二个观点的早期支持者赞成量子力学的“隐变量理论”(hidden variable theory)。但近期文献却偏向于支持第一个观点：物质世界在本质上是充满或然性的。我们都在期待该理论中的很多悖论能够被解决，并试图知晓该理论对物质世界本质的意义。虽然物理学中的争论已经被用来证明社会科学哲学基础的合理性，但这些争论并不影响社会科学中的研究实践及推论逻辑。

够在所描述的对象中区分出系统性方面与特异性的(idiosyncratic)方面。因此,上文关于描述性推论的主张不是去质疑案例研究式的历史研究。任何社会科学推论都应该满足本书讨论的基本原则。即使从证据中寻找系统特征非常困难,也绝不能忽视这些原则。

为了便于读者理解,让我们举个历史研究中描述性推论的例子。假设研究者试图了解 1955 年至 1990 年间美苏首脑峰会的结果。该研究的最终目标是要回答如下因果问题:在什么条件下以及在多大程度上,峰会增进了两国合作?回答该问题需要解决很多因果分析中的难题,尤其是那些涉及系统性变量之间因果关系方向的问题。[①]不过在本节中,我们只讨论和描述性推论有关的问题。

假设通过历史分析、调查专家们的意见、搜集“合作”与“冲突”事件的数量或者将所有这些信息结合起来去评估两个超级大国的峰会在多大程度上会促进它们间的合作。对于促进合作的因素,比如权力更迭、美国的选举周期、两国的经济状况、双方预期已经实现多少,等等,都可以设置为有待检验的假设。同时假设要解释的是两国每年的合作水平:试图将它与前一阶段两国是否举行过首脑会晤以及其他控制变量联系起来。

研究者观察到的只是每年实际发生的合作程度(即使关于合作的测量指标是完美的)。即使在两国峰会召开之后的几年观察到高水平的合作,但在没有进一步研究的情况下,研究者还是不能确定峰会与接下来的合作是否系统相关。如果观察值很少,峰会与合作间的相关性也许反映的是根本性的不确定性(观点 1)所导致的随机性,当然也可能是由那些尚未识别出来的解释变量(观点 2)导致的随机性。那些尚未识别的解释变量可能包括导致苏联作物歉收的气候变化、军事制衡的破坏、领导人更迭,所有这些因素都会导致合作程度的改变。如果可以搜集到这些遗漏变量的信息并由此识别其作用的话,就有可能作为可被检验的替代性解释。如果无法识别其作用,这些变量就可被当作

---

① 我们把这个问题称为“内生性”(endogeneity),该概念将在本书 3.5 节中被讨论。对合作的预期可能会反过来影响峰会的召开。在这种情况下是合作解释了峰会,而非峰会解释了合作。如果行动者是理性的话,则这种逻辑不足为奇。

影响两国合作程度的非系统性因素了。为了排除这种情况,研究者就需要搜集其他年份的信息。由于随机事件是不可持续的,因此这些因素不可能在不同年份都导致合作程度的差异。因此,我们建议读者:只有在不同背景下(在这个例子中就是不同年份)进行重复检验,才能判断其结果到底是由系统性的因素导致还是仅仅由随机暂时性的因素导致。

对于研究者来说,区分去系统过程与非系统过程并不容易。从社会科学的角度来看,一场影响工薪阶层胜过中产阶级的流感对 1979 年假设重复选举而言是一个不可预测的(非系统的)因素,但可能降低工党获得的选票。工党的失败只是 1979 年重复选举试验中的一个结果。但是,疾病袭击呈现出的阶级差别可能成为多次重复试验中减少工党选票的系统性因素。

由于性格或在电视辩论中失误,一位候选人在美国大选中战胜了他的对手,在冷战时期这可能成为一个影响美苏合作的随机性因素。但是,如果对选民有感染力的竞选口号可以缓和与苏联的关系,那么主张怀柔政策候选人的获胜就构成了解释双边合作的系统性因素。

系统性因素的影响是持久的,且当取一个特定值的时候会导致一致的结果。非系统性因素则是暂时的,研究者无法预测他们的影响。但是,这并不意味着系统性因素是不变的。竞选口号可能是解释选举行为的一个系统性因素,但这并不意味着口号本身是不变的,不变的是竞选口号对选举结果的影响。即使该作用可变,也是以一种可以被预测的方式变化。比如,当美苏关系缓和的时候,怀柔政策有助于在选举中赢得选票;当关系紧张时却会有相反的结果。类似地,气候状况可能是随机因素(比如间歇性且不可预测的气候变化导致的不可预测结果),也可能是系统性因素(如果坏天气总是导致那些主张怀柔政策的候选人获得更少的选票)。

简单地讲,概括历史细节是使用数据过程中一个重要的中间步骤,同时研究者也不能忽略可以区分处随机性与系统性现象的描述性推论。仅仅了解给定环境下到底发生了什么是不够的。如果不将其中的系统性特征区分出来,历史的教训就会荡然无存。对研究对象中哪些

方面会持续下去，哪些方面对未来事件和研究有启发意义，我们也将一无所知。

2.7 判断描述性推论的标准

在本章的最后一节，我们将介绍统计学中被普遍使用的评价推论的三个标准：无偏性（unbiasedness）、有效性（efficiency）及一致性（consistency）。针对每个标准的讨论都以本章 2.6 节介绍的随机变量框架作为基础。这些标准对验证和改善定性研究有直接和显著的作用。为了弄清楚这些概念，本部分只使用最简单的例子，且这些例子都和描述性推论相关。简单地说，推论含有对参数的估计，包括估计描述性推论中随机变量的期望值或方差（μ 或 σ^2）。我们也用同样的标准去评判因果推论（见本书 3.4 节）。本章剩余部分将集中讨论这些概念，而这些标准对定性研究的具体建议则放到之后的章节讨论。

2.7.1 推论的无偏性

如果研究者反复应用同一种推论方法，得到的估计值有时会太大，有时则太小。平均来说，进行大量尝试后能否得到正确的估计呢？如果答案是肯定的，那么这种方法或者“统计数据”就被认为是无偏的。统计数据的这个性质对每一次推论结果偏离平均值有多远并没有要求，只要平均起来是正确的就可以了。

如果测量偏误是非系统的，这意味着该偏误有时偏大有时偏小，那么估计结果就是无偏的。在一系列重复试验中，如果存在系统的测量误差而导致估计结果更多地偏向一个方向，估计结果就是有偏的。在对约旦河西岸社区中冲突程度的研究中，假设当地领导人刻意制造冲突以影响研究的结果（他们这样做也许出于某种利己的政治企图）。平均来说，研究者在每一个社区观察到的冲突程度都是有偏的，一般而言

要高于真实冲突水平。另外一个例子,假设1979年的选举都在星期日举行(其实可以在任何一天举行),如果这一事实可以系统性地帮助其中一方(比如,保守派出于宗教原因不愿意在星期日投票),估计结果也会存在偏差。再或者选票计票人员中存在腐败行为,不公正地偏袒某一方,基于他们报告获得的估计结果也将是带有偏差的。然而,如果选举在不同日子举行且日期的选择与研究者感兴趣的变量无关,那么测量结果就不会因为选举举行在特定某一天或者某一党派受到偏袒而存在偏误。最后,假如由于选票计数人员随机的粗心而导致票数计算错误,其估计结果仍然是无偏的。

假设依据法律,英国选举一直在星期日举行或者那种偏袒一方的计票方法长期存在(也许因为某种特殊的投票方案或是持续的腐败行为),一旦包含这些系统性特征,研究者也是可以预测到那种围绕平均选票数变化的估计值的。由此可见,偏差并不仅仅存在于数据之中,还取决于那些被研究的理论。由此可见,仅仅声称数据是有偏的并没有多大意义,即使这个数据可能真有许多错误。

在这个例子中,研究者就需要从选举制度中存在的“实质性的偏差”(substantive bias)去区分估计值中的“统计偏差”(statistical bias)。关于后者的例子是,投票时间的选择使得工薪阶层难以参选,这是选举制度中很常见的实质性偏差。研究者希望能估计到实际选举(含有实质性偏差)中的平均选票数,也希望估计出没有实质性偏差的假设选举下的平均选票数,该假设假定选举是不存在投票时间上的限制的。这样做就使我们估计出系统中实质性偏差的大小了。不论通过哪种方式,研究者希望得到的是一个统计学上无偏的估计值。

社会科学研究用到的数据非常容易受到偏差的影响,这种偏差有些来自那些提供原始资料给研究者进行描述性推论的人。他们经常出于某些目的过高或过低地估计取值,对此研究者应该保持警惕的态度。政府机构有动机去高估那些新项目的效果,以此获得更多资金支持,同时他们也有动机去低估失业率来证明他们忠于职守且成效卓著。改善这种有偏的估计值需要更深入的研究。比如,在迈伦·韦纳(Myron Weiner, 1991)关于印度教育和童工的定性研究中,为了解释在义务教

育施行方面印度为何表现得比很多其他国家都差，他必须首先弄清楚印度的入学率到底有多低。从印度某个邦的官方统计数据中，他发现适龄儿童的入学率是98%。通过更深入的调查后，他发现这么高的入学率是在儿童刚入学的时候统计的，而且只统计这么一次。之后这些儿童就被想当然地认为在学校里待满了七年，哪怕他们在学校的时间只是一天！这种仔细的调查显示出实际入学率比官方吹嘘的数字要低很多。

关于无偏性的公式表述

假设要估计公式(2.2)中的 μ，并用平均值作为估计值，$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$。在某个数据中，$\bar{y}$ 是在所有 $n=650$ 个选区中工党所获选票的平均数(或者是约旦河西岸所有社区中的平均冲突水平)。如果考虑每个选区进行的选举可以无限次重复试验，样本均值就变成了650个随机变量的函数，表示为：$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$。这样，样本均值就相应地变成了一个随机变量。对于这些假设上的重复试验，$\bar{Y}$ 会近似于 μ；而对于其他试验，它们之间可能会相差甚远。问题是通过无数次的重复试验，平均起来 $\bar{Y}$ 是否与 μ 相等？为了找到答案，我们再一次运用期望运算，因为该运算可以获得无限次重复试验之下的平均值。期望运算的过程如下：

$$E(\bar{Y})=E\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(Y_i)=\frac{1}{n}n\mu=\mu \tag{2.4}$$

从上述公式可以得知，$\bar{Y}$ 是 μ 的无偏无计量(与统计课本相比，这个例子可能不够正规，但是主要特征相同)。

2.7.2 推论的有效性

我们很少有机会将估计值的计算应用到很多场合中，即使这些场合在本质上相同。事实上，除了一些设计精巧的实验，大多数情况下研究者只有一次机会。在这种情况下，无偏性虽然是我们感兴趣的指标，但研究者想确认该估计值在多大程度上接近真实值。有效性概念就有助于在无偏估计值中进行进一步区分，也有助于在那些带有少许偏差的估计值中进行区分（在评估有效性之前，那些偏差过大的估计值就应该被排除掉）。

有效性是一个相对概念，它通过计算重复试验中估计值的方差来估测。对于无偏估计值，它的方差越小意味着估计的有效性就越高，因为小的方差表示估计值与真实参数值更加接近。而对一个偏差过大的估计值，就没有必要关注它的有效性了。在这种情况下，即使是小方差也不可能使估计值接近于真实值（实际情况是，它们紧密地分布在错误值周围）。因此，我们只关注偏差较小时的有效性，当然在有些情况下也可以牺牲一点无偏性去换取更高的有效性。

如果我们试图估计西岸社区中巴以冲突的平均水平，有两种估计方法可供选择：第一种是从某社区中选取单一且典型的观察值，另外一个方法是从 25 个社区中都选出相似的观察值。显而易见，后者要好过前者，当然前提是研究者在每个社区投入的努力都和第一种方法中单一社区投入的努力一样多。这就告诉我们为何要尽可能多地去获取可以验证理论的现象，同时也有利于研究者在多种数据组合中作出最优选择。

有效性使得研究者可以比较从单一观察值 ($n=1$)的案例研究和从大样本 ($n=25$)研究中获得的估计量。后者是指针对西岸 25 个社区分别进行为时一周的研究后得到的平均冲突水平。如果应用合理，两个估计值都可能是无偏的。运用同样的模型，单一观察值估计值的方差是 $V(Y_{typical})=\sigma^2$。这意味着，研究者本想要选择那些我们认为是“典型”的地区，事实上这个地区同样受到随机变量的影响。在大样本

研究中，估计值的方差是 $V(\overline{Y})=\sigma^2/25$，也就是样本均值的方差。可见，前者的大小是后者的 25 倍(即有效性更低)。因此，通过这个例子读者就能理解为何观察值越多越好了。

读者需要注意的是，详尽的个案研究在某些条件下也可以获得与大样本研究一样甚至更好的结果。确切地说，尽管研究者都应该尽可能多地收集观察值(只要这些资料是研究所需)，但在某些情形下，与那些急于搜集更多观察值但每一个都不够具体和确定的研究相比，深入而详尽的个案研究可能更加有效。

如果所有条件都相同，观察值才是越多越好：因为随着观察值的增多，差异性(即较低的有效性)就会降低。一致性所描述的就是当观察值数量变得非常大时，变异性趋近于零的性质。此时估计值就与我们试图估计的参数相等了。①

但是在实际情况中，并非所有条件都是相同的。例如，如果某些因素影响对研究对象的测量，就有可能导致测量值大幅偏离真实值(即估计值的方差很大)。当然，在这种情况下，研究者可以通过其他研究途径来了解这些因素到底是什么。比如说，可以进一步假设观察这些因素和消除它们影响的能力是随研究社区数量的增加而大幅度下降(针对如此多的社区，出现该现象的原因可能是研究者缺乏足够的时间和知识来消除这些因素的影响)。此时，研究者就必须权衡到底选择包含很多观测值的单一案例研究，还是所有 25 个案例的一揽子研究(每个案例含有一个观察值)。

如果一个案例研究仅仅包含一个观察值，那么包含 25 个观察值的研究优势就很明显。不过案例研究的优势在于它更易于被读者所理解(如果可以用公式化表示的话)。研究者可以谨慎地先选择一个社区，使它尽量具有代表性，或通过它读者可以理解该社区与其他社区间的关系。对于前者，通过询问当地居民或阅读当地报纸可以知道这个社

① 读者需要注意的是，一个无偏估计值也可能是不一致的。比如，Y_1 是 μ 的一个无偏估计值，但却不是 μ 的一致估计值。因为随着样本的增大，该估计值的估计效率并不会改善(事实上根本不会发生任何变化)。而有些估计值虽然是有偏的，却是一致的。比如，$\overline{Y}-5/n$ 是个有偏的估计值，该值却是一致的：因为当 n 趋于无穷大时，$5/n$ 趋于零。

区是否具有代表性，尽量避免由于某些非系统性因素导致该社区被选中。对于后者，研究者就需要调整观察到的冲突水平，使之和约旦河西岸社区中冲突的水平(μ)相近。这是案例研究估计工作中最困难的部分，因此要非常谨慎地提防偏差乘虚而入。当确信偏差已经被最小化后，研究者便可以将注意力转到增进有效性上来。要实现有效性的改进，研究者需要花费数周时间在社区中进行大量的研究。这些研究包括采访社区领袖、普通居民以及学校教师，与孩子们交谈，阅读当地报纸，追踪家庭的日常生活，等等。通过这些步骤，在该社区中获得的信息将远远超过 25 个观察值，同时也可以得到一个不存在偏差的案例研究。这样做比那种研究所有 25 个社区来得更有效率。

再来看另外一个例子。假设研究者正在进行一项"国际毒品问题"的研究，希望测量某个区域内可卡因种植面积占当地所有农业用地的比重。有两种研究方案可供选择：第一种是基于单一村落的案例研究，另外一种是对该区域所有村落进行大规模的统计学研究。乍看起来，研究者也许觉得研究整个区域的方法会更好，只是要这样研究会受到很多现实因素的限制，研究者必须使用地方政府提供给联合国专门机构的数据。众所周知，这些数据是很难反映真实情况的，因为它们是该国外交部基于公共关系的考虑而精心准备的。前提是我们可以访问并近距离地观察一个村落，从而有可能更正政府提供的官方数据，使之更接近真实值。那么该选择哪种研究方法呢？只研究一个村落，还是研究两三个村落？我们更倾向于集中对一个村落进行研究，并用得到的研究结果去解释和完善政府提供的关于其他村落的数据。因此，到底选择哪种方法取决于哪组数据能最好地回答研究问题。

再看一个例子，假设我们着手研究欧洲共同体，试图估算由委员会和部长会议制定的那些针对共同体中某一产业的管制规则的严格程度。研究者首先要收集该产业部门已经施行的那些管制条例，然后根据这些条例的严格程度编码，最后估算出整体的平均严格程度。如果能获得 100 个具有相似严格程度的规则，那么测量的方差就是任一给定规则的方差除以 100($\sigma^2/100$)。如果规则之间具有相关性，得到的

方差会更小。毫无疑问,相对于仅仅考察一个规则并将它作为整个产业规则严格性的估计值的方法,上述测量结果要好得多。

实现上述方法的前提是正式规则字面上的严格程度等于实施过程中的实际严格程度。如果进一步研究这些规则在实际情况下的执行情况,研究者也许会发现名义上的规则与实际情况出入很大。这样一来,对规则名义上的测量可能存在严重的系统性偏差:倾向于高估规制的严格程度。在这种情况下,我们又一次面临在偏差与效率之间作出权衡。那么对规则实际的执行状况进行三四个深入的案例研究以获得名义和实际之间的关系就非常有必要。研究者可以用基于三四个偏差较小但有效性也较低的案例得出估计值,以此替代所有 100 个案例的估计值。如果这种方法可行,通过对这几个案例的深入研究来校正从 100 个案例中获得的指标中存在的偏差,然后用校正过的指标作为最终估计值将更加有效。在这个过程中,深入的案例研究就与大样本技术结合在一起了。我们认为这种方法在目前的社会科学研究中应当被更加频繁地采用。

那些深入研究某一领域的学者对案例研究做的辩护就隐含在上文的例子中。那些依靠数字的大型研究工作对于那些初级数据研究人员来讲并不容易上手(这些人可能没有注意到某地特殊的选票统计方式,而错误地假设它们与选民有某种联系)。只有那些收集一手资料并了解资料来源的研究者才能对其进行修正。在接下来的部分,我们将阐述怎样才能更系统地做出选择。

下面是对这种选择有效性的公式化分析。该分析以英国选举为例,告诉读者如何做出取舍。在遇到其他具体例子时,这些分析中蕴含的逻辑将有助于大家做出更好的决定。当然,这些决定也始终要求读者具备定性判断的能力。

关于权衡有效性的公式表述

将样本均值 $\bar{Y}$ 的方差记为 $V(\bar{Y})$。在简单随机抽样中,随机变量方差的计算公式为:

$$V(\overline{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i)$$

进一步假设选举在各个地区假设性的重复试验中方差都是 σ^2，那么由此获得样本均值的方差：

$$V(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) \tag{2.5}$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$$
$$= \frac{1}{n^2} n\sigma^2$$
$$= \sigma^2/n$$

在这个例子中，$n=650$，因此大样本估计量的方差是 $\sigma^2/650$，而在案例研究中的方差是 σ^2。除非能通过定性的、随机误差的校正方法把后者减小至少 650 倍，否则出于有效性的考虑，我们自然更偏向于统计学估计。

最后，对于如何在偏差和有效性之间做出选择，我们也应当给予足够的重视。在一个由无偏观察值组成的数据中，即便仅用前两个观察值求得均值，该均值与全部观察值的样本均值也一样是无偏的。不过，这样做却浪费掉了太多信息：虽然不改变无偏性，却大大降低了估计的有效性。因此，如果没有一个关于有效性的标准，读者就不知道到底选择哪一个估计值。

如果我们想了解民主党在下一届总统选举中能否获胜，因此随机抽取了 20 个美国籍的成年人，问他们打算将选票投给哪个党派（在这个随机抽样的简化版本中，假设我们从所有成年美国人中抽取调查对象，同时每个人被选中的概率也相同）。假设另外有个研究者以 1 000 个居民作为研究对象进行相似的调查。此时是否应将这

1 020 位受访者合并起来然后进行估计呢？假如这 1 000 个观察值也与之前的 10 个一样是通过随机抽样获得的，那么就应该将他们并入我们的研究：合并以后获得的估计量仍然是无偏的，有效性却会大大增加。

但是，假设这 1 000 人中只有 990 个人是随机抽取获得，另外 10 个人则是民主党议员，他们是在随机抽样之后意外被纳入到数据中的。进一步假设研究者无法区分出这 10 个观察值和那些随机抽取的观察值，因此没法将它们移除出去。这样一来，基于这 1 020 个观察值的估计值将导致对民主党胜选可能性的略微高估。因此，包含这 1 000 个额外观察值会导致对总体估计值的偏差，不过却可以大幅度地增加估计的有效性。要不要包含这 10 个观察值取就决于在有效性方面的改进是否能够弥补在偏差上的扩大。从直觉上来说，与仅依靠 20 个观察值做出的估计相比，基于 1 020 个观察值做出的估计值更可能接近真实值。尽管在实践中可以同时采用这两个方法，我们仍然偏向于采用基于大样本的估计量，只要它引入的偏差足够小（在这个例子中，我们是可以判断偏差的方向的，因此能够进行相应的校正）。

如果有充足的定量数据，研究者就可以将上述问题用公式化表述出来以便做出最终决定。尽管研究中存在的定性特点使得权衡很困难甚至不可能，但对它的理解仍然有助于获得更加可靠的推论。

偏差与有效性对比的公式化表述

现在考虑两项估计工作，一项是带有偏见的研究者基于大样本进行的估计，另外一项是一个公正的研究者进行的无偏但有效性较低的小样本研究。假设我们希望得到 μ 的估计值，基于大样本研究得到的估计值为 d：

$$d = \left(\frac{1}{n}\sum_{i=1}^{n} Y_i - 0.01\right)$$

我们使用 μ 的另一个估计值 c，在小样本的研究中可表示为：

$$c=\left(\frac{Y_1+Y_2}{2}\right)$$

其中，地区 1 和地区 2 都是有代表性的选区，所以有 $E(Y_1)=\mu$，且 $E(Y_2)=\mu$。

至此，研究者应该选择哪一个估计值呢？第一个答案是哪个都不选，而是采用样本均值 $\bar{y}$；也就是那个由公正的研究者基于大样本研究得到的无偏估计值。当然，最好或最明确的估计值是不存在的。为了回答这个问题，我们必须对基于这两个方法获得的估计值的偏差与有效性进行评估。

首先评估偏差，通过简单的运算，可以得到第一个估计值 d 是稍微带有偏差的：

$$\begin{aligned}E(d)&=E\left(\frac{1}{n}\sum_{i=1}^{n}Y_i-0.01\right)\\&=E\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)-E(0.01)\\&=\mu-0.01\end{aligned}$$

运用相似的运算规则，可以证明第二个估计值 c 是无偏的：

$$\begin{aligned}E(c)&=E\left(\frac{Y_1+Y_2}{2}\right)\\&=\frac{E(Y_1)+E(Y_2)}{2}\\&=\frac{\mu+\mu}{2}\\&=\mu\end{aligned}$$

根据以上计算，研究者会选择估计值 c，也就是那个来自公正研究者的小样本研究，因为它是无偏的。平均来说，在无限次的重复试验中，如果调查者是有偏见的，哪怕偏见程度很轻微，d 也是错误的。和它相比，估计量 c 却在平均水平上是正确。

不过根据有效性标准做出的选择就不一样了。首先计算各个估计值的方差：

$$
\begin{aligned}
V(d) &= V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - 0.01\right) \\
&= V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - V(0.01) \\
&= \sigma^2/n \\
&= \sigma^2/650
\end{aligned}
$$

该方差和样本均值的方差相等，因为 0.01 在样本中不会变化(这意味着方差为零)。类似地，我们计算出 c 的方差①：

$$
\begin{aligned}
V(c) &= V\left(\frac{Y_1 + Y_2}{2}\right) \\
&= \frac{1}{4}[V(Y_1) + V(Y_2)] \\
&= \frac{1}{4} 2\sigma^2 \\
&= \sigma^2/2
\end{aligned}
$$

因此，$V(c)=\sigma^2/2$ 比 $V(d)=\sigma^2/650$ 大出 325 倍，所以，估计量 c 在有效性方面要低很多，该结果也是符合直觉的，因为数据中的大多数信息都没有被 c 用到。

至此，读者该选择哪一个呢？估计量 d 虽然有偏，但与 c 相比却更加有效；估计量 c 虽然无偏却缺乏有效性。在这个例子中，我们选择倾向于 d。鉴于偏差相当小(0.01)，因此宁愿牺牲一点无偏性去换取在有效性方面的显著改进。但在其他例子中，有效性的改进并不足以弥补偏差，在那种情形下，估计值也许远远偏离真实值。权衡偏差—有效性的公式化指标是计算均方误差(mean square error, MSE)，该指标是偏差与有效性的结合。假设 g 是参数 γ(希腊字母 Gamma)的一个估计值，MSE 的定义如下：

---

① 在计算这两个估计值的方差时，下文表达式的第二行中假设选区之间并不存在空间上的相关性。

$$MSE(g) = V(g) + E(g-\gamma)^2 \tag{2.6}$$
$$= \text{variance} + \text{Squared bias}$$

由此可见，MSE 是方差与偏差平方的相加(参见 Johnston，1984:27—28)。这表明研究者应该选择具有最小均方误的估计值，因为该指标告诉我们那些有偏但方差更小的估计值比较好。

上文例子中的均方误差计算如下：

$$MSE(d) = \frac{\sigma^2}{650} + (0.01)^2 \tag{2.7}$$
$$= \frac{\sigma^2}{650} + 0.000\,1$$

以及

$$MSE(c) = \frac{\sigma^2}{2} \tag{2.8}$$

对于 σ^2 的大多数取值来说，都有 $MSE(d) < MSE(c)$，所以相对于 c，我们更倾向于选择估计量 d。

从理论上来说，我们倾向于无偏估计值，同时也希望它的有效性越高越好。在接下来的一章中，我们将告诉读者在实际研究中对偏差和有效性做出权衡是十分重要的。

3 因果关系与因果推论

我们已经在第2章中讨论了社会科学研究的两个阶段：归纳历史细节（第2.5节）与描述性推论（第2.6节），其中描述性推论将研究对象划分为系统和非系统两部分。然而，许多针对社会和政治现象的研究却到此为止。这些研究似乎刻意地回避作因果推论，它们的研究者更希望手头掌握的事实能够"不言自明"。

和历史学家一样，社会科学家也需要对历史细节进行归纳并作出描述性推论。尽管如此，对于社会科学来说，没有因果推论的研究往往是不完整的。换句话说，精准的描述性推论虽然可以作为因果推论的前提，但是研究不能只有描述性推论。当然，这并不是说所有的社会科学研究必须都作因果解释。在因果推论比较困难的情况下，描述性推论也可以作为研究的最终目的。

基于上述原因，研究者必须先明确其研究目的是去描述还是作解释。因果推论往往会让许多社会科学家困扰：他们牢记"相关关系并非因果关系"这条戒律，对因果假设和因果推论抱有戒心，甚至直接声称自己的研究是"对相关关系而非因果关系的研究"。当然也有人走到了另外一个极端，一些研究者随心所欲地进行因果陈述，将建立在不明确的研究设计之上且未经证实的假设或推测都称为"解释"。①其实以上

① 由于社会科学家更倾向作解释而不是"仅仅描述"，因此一些学者在研究较复杂的对象时经常用解释性术语来点缀。如果不这样做，他们害怕自己的研究就会被别人认为是二流水平。解释工作应当建立在因果推论之上。文献中"非因果的解释"这样的表述确实很让人费解，其实质反映了关于因果解释及内部不一致性的争议。因此，如果研究者没有做出解释，其原因并不在于研究能力和想象力的缺乏，主要是因为问题本身的难度及重要性。对此，研究者不必感到不安。读者需要注意，在那些重要的研究工作中，正确的描述比拙劣的解释有价值。

这两种情况都是在回避因果推论问题。

如果一个研究目的是作因果推论却缺乏相关的描述，那么这个研究就不符合科学推论的原则从而有点不着调。虽然研究者不可能消除因果推论中的不确定性，但这种不确定性本身不能阻止我们去尝试因果推论的工作。研究者应该在条件许可的时候进行因果推论，同时如实表述该推论的不确定性。只要能严谨地处理不确定性，就可以大胆地进行因果推论。在具体的研究过程中，所作的因果假设应该尽可能地符合因果推论的原则，这点非常关键。本书第 4 至 6 章中的大部分内容将详细说明因果推论的适用情况及在这些情况下定性研究者如何为其因果假设提供可靠的证据。

本章的 3.1 节将定义出定性和定量研究中的因果关系。3.2 节将廓清已有文献关于因果关系的不同概念。需要说明的是，这些概念在本质上与本书的定义并不冲突。3.3 节则讨论关于现实世界及有效因果推论所要求的假设。3.4 节将告诉读者如何把描述性推论中得到的判定标准应用于因果推论。最后在 3.5 节中，就如何构建因果解释、因果理论和因果假设，作者将给出一些建议并以此总结全章内容。

3.1 定义因果关系

在本节中，我们先将因果关系定义为独立于数据的一个理论概念，然后结合数据理顺因果推论的逻辑（针对因果推论中具体问题的讨论，见本书第 4 至 6 章）。其中第 3.1.1 节通过一个定量例子给出因果关系的详细定义，在接下来的 3.1.2 节中，我们再用一个更为复杂的定性例子进一步诠释该关系。

3.1.1 定义及一个关于定量研究的例子

本书定义的因果关系可以简单明了地适用于基于一个单位的研

究。①正如 2.4 节给出的定义，所谓的一个单位是指在研究中观察到的诸多要素中的一个，例如一个人、一个国家、一年或者是一个政治团体。为了让读者更好地理解，我们举一个定量研究的例子：美国民主党候选人在众议院中是否有议席对其在选战中得票比例的作用（只研究民主党是为了表述方便）。在这个关系中，被解释变量是民主党在众议院两党选举中得票的比例。相应地，关键解释变量就有两个值：民主党人在众议院中已经有议席还是没有（为了进一步简化，我们只考虑共和党最终落败的选区）。

关于因果的术语有很多表述方式，在以下的表述中都会涉及。一般来说，“被解释变量”也被称作“结果变量”（outcome variable）；“解释变量”（explanatory variable）常被称为“自变量”（independent variable）；我们将解释变量进一步区分为“关键因果变量”（key causal variable）（也称作“原因”[cause]或“介入变量”[treatment variable]）和“控制变量”（control variable）。最后，关键因果变量经常会取两个或两个以上的值，一般用“介入组”（treatment group）和“控制组”（control group）表示。

为了方便读者理解，让我们仅考虑 1998 年纽约第四国会选区的情况。该年有一个民主党现任议员候选人和一个共和党非现任议员候选人角逐国会议员席位。假设民主党候选人在这次选举中获得的选票比例是 y_4^I（其中，右下角的数字 4 表示选区的编号：纽约第四国会选区，右上角的字母 I 表示该候选人是现任国会议员），这样，y_4^I 就是被解释变量的取值之一。为了定义出因果作用（理论上的作用），让我们假设民主党的这个现任议员在选举的一开始就决定放弃参选，因此民主党提名另一个非现任的候选人（此人是初选的获胜者），除此之外其他一

① 本小节内容的出发点源于霍兰（Holland，1986）对于因果关系的研究，他在大量研究的基础上给出了因果关系的定义，并称之为“鲁宾模型”。该定义与唐纳德·鲁宾（Donald Rubin，1974，1978）的研究紧密相关，同时也从亚里士多德、洛克、休谟、穆勒、苏佩斯、格兰杰、费希尔及内曼等人的研究中获取了灵感。在本书中，我们基于苏佩斯等学者剔除的“或然性因果关系”（probabilistic causality）来进一步拓展霍兰的这个定义。我们认为这种拓展是必要的，因为迄今为止还没有一个方法在定义因果关系的同时，将该因果效应区分出系统和非系统两个部分。

切都没有变化。对于这个非现任的民主党候选人会获得的选票比例我们用 y_4^N 来表示(这里 N 表示该候选人不是目前的国会议员)。[①]

上述定义出来的反事实对探索因果关系非常重要,实际获得的选票比例 y_4^I 和假设条件下获得的选票比例 y_4^N 间的差异就是真实的因果效应。在接下来的内容中,我们将进一步描述这个概念以便于读者理解。需要注意的是,尽管反事实明显与事实相悖,但对它的定义必须非常小心谨慎。其次,反事实必须是合理的并且在精确设定的条件下是可能发生的。定义反事实的关键在于改变介入变量的取值时,其他影响因素需要保持不变。就上面这个例子来说,关键因果变量(或者叫"介入变量")是参选者目前在国会的任职情况:从"现任议员"转变为"非现任议员"。这种转变是假设出来的,但除此以外,研究者必须要在民主党做出提名决议的同时保持其他所有条件不变。这些条件包括该区以往竞选中民主党和共和党相对实力的对比、选举提名程序、选区特点以及当时的经济及政治氛围,等等。不过我们不会去控制候选人的特质,比如知名度、曝光率、对国会运作的了解程度以及其他与提名相关的因素。不作上述控制的原因在于,这些特质正是介入变量现任与否所产生的部分结果。换句话说,现任议员的优势当然包括较高的知名度和曝光率。如果连这些因素都被控制住,那么现任与否产生的一些重要影响就无法被识别出,最终导致误判其在选举过程中发挥的作用。事实上,对上述这些特质的过度控制会导致研究者错误地认为现任与否根本不影响选举结果。[②]

更正式地说,是否为现任议员在纽约第四选区所产生的因果效应在

① 关于该例子的详细讨论见格尔曼和金(Gelman & King, 1990)的研究。一般而言,I 和 N 分别表示"处理/介入"和"控制"组,也可以表示在任何实际或理论上所进行的两组区分和处理。当然,是否将解释变量的某一取值称为"处理/介入",而将另外一个取值称为"控制",研究者可以任意决定。

② 乔恩·埃尔斯特(Jon Elster)认为在很多情况下"因果关系不能通过反事实来陈述"(Jon Elster, 1983:34—36)。比如,实际生活中,一些因素会同时影响解释变量和被解释变量。在我们看来,埃尔斯特所指的仅仅是推论中存在的普遍问题,而推论总是有一定的不确定性的。但这些困难不能作为反对用反事实陈述因果关系的理由。尽管如此,埃尔斯特仍然意识到反事实推论"在因果分析中扮演着重要角色"(1983:36)。因此,与其认为埃尔斯特反对采用反事实进行推论,不如将他的观点视作对该方法的批评。

这个研究设定中就是民主党候选人是否为现任国会议员所影响的选票比例。该因果效应是现任与否在获得选票比例上的差别，即($y_4^I - y_4^N$)。为了表述的清晰，我们将这种差别称为“实现了的因果效应”(realized casual effect)，在下面的公式中用 i 代替第四选区使得结果更具一般意义①：

$$\text{在第 } i \text{ 个单位中实现了的因果效应} = y_i^I - y_i^N \tag{3.1}$$

当然，上述因果效应仅是理论上的定义，在实际进行的选举中两者绝不可能同时出现：我们要么只能观察到 y_4^I 和 y_4^N 中的一个，要么两者都看不到。因此，从这个简单的定义中我们就能知道研究者是无法获得确实的因果效应的。霍兰将这个问题称为“因果推论的根本问题”(the foundational problem of causal inference)(1986)。这个问题确实称得上是根本问题，因为不论研究设计有多么完美，收集到的数据有多么丰富，研究者的洞察力有多么敏锐，也无论研究助理如何勤奋，实验控制得多么精确，我们永远都无法获得一个确定的因果推论。实际上，本书讨论的研究设计涉及的大部分实证问题都与该根本问题相关，我们给出的所有建议也都围绕如何避免该问题展开。

本书对因果关系的定义和霍兰的有所不同。在 2.6 节，我们说过社会科学经常将世界划分为系统和非系统两部分，而霍兰的定义并没有区分这两个部分。②为了进一步说明这种区分的重要性，设想我们重新举行 1998 年纽约第四选区的选举。如果选战是在一个民主党现任议员候选人和一个共和党候选人之间展开，结果会是什么样子呢？即使竞选的基本条件相同，但由于竞选中存在一些非系统因素，这些因素往往在每次竞选中都有些许差别，这就会导致投票结果产生微小的改

① 在公式(3.1)中，我们可以把“i”替换为“4”以特指研究中的第四选区。

② 霍兰之所以没有进行这种区分，也许是因为作为一名统计学家，他更倾向于接受本书 2.6 节“观点 2”中的内容。针对因果推论中的根本性问题，他给出的“统计解决方案”与本书对因果效应的定义极为类似。但他的定义没有将因果推论局限在一个研究单位，而是通过观察更多单位来解决这一根本问题。对于霍兰来说，期望值是每个研究单位的平均值(Holland, 1986:947)，而本书的计算方法(见下文)则是通过在同一单位上重复进行相同的试验获得。

变。这些非系统因素可能包括竞选演说中的口误、对某一问题的演讲及所持立场获得了超乎寻常的支持率、在辩论中出丑、巡回演讲或竞选当天突遭恶劣天气或者狗仔队忽然捅出的花边新闻。

我们假设这次选举可以重复进行，对于每次重复选举中民主党候选人获得的选票比例用一个变量表示。在2.6节已经提到这个变量叫做“随机变量”，之所以取这个名字是因为该变量要么含有一些非系统特征，这些特征虽然受解释变量影响但并没有在理论考察范围内，要么它根本就是一些不可解释的因素。①我们用 Y_4^I 来定义被解释变量（注意是大写字母 Y），即现任议员民主党候选人获得的选票比例，相应地用 Y_4^N 表示非现任民主党候选人在同样条件下获得的选票比例。

现在将该选区的随机因果效应（random causal effect）定义成这两个随机变量间的差异。为了公式表述的一般性，再次将代表第四选区的符号4转换成单位 i：

$$\text{对于单位 } i \text{ 的随机因果效应} = Y_i^I - Y_i^N \tag{3.2}$$

（和随机变量的定义类似，随机因果效应就是从不断重复的相同实验中获得的因果效应。该因果效应可以呈现出竞选的诸多系统性特征。）假设可以同时观察到第四选区在两种情况下的选票比例——一个是现任议员的民主党候选人，另一个为非现任候选人——我们就能够直接观察到公式(3.1)所表示的“实现了的因果效应”。可是由于刚才提到的因果推论中的根本问题，我们实际上无法看到这样的因果效应。因此，公式(3.1)“实现了的因果效应”其实是无法观测到的，因此只能通过公式(3.2)中的随机因果效应表示。换句话说，如果我们可以在第四选区现任和非现任候选人之间重复进行相同选举试验，其“实现了的（实际

① 正如本书2.2节详细解释的，该表述会令人困扰。“随机变量”中也包含系统性的部分，因此并非总是不可预测的。该术语在统计学上有其特定含义，其背后体现的概念也比较重要。在统计学中，随机性并不意味着“怎么都行”或者“什么时间都能发生”，它表示的是很多可能的概率过程中的一个。例如，决定硬币哪面着地的随机过程在本质上和影响欧洲经济共同体日益官僚化的随机过程是不同的，和那些决定意大利选举制度变革所导致的不确定性政治后果的随机过程也是不一样的。我们表述的关键在于，所有这些“随机”事件都是由“系统性”和“随机性”两部分构成的。

上无法观测到)因果效应"就是随机因果效应。

将因果关系描述成随机变量的一个系统性特征也许会使问题变得有些复杂。不过这样做有两个好处:第一,它可以将因果关系类比成某个现象的系统性特征(如均值或方差)。而这正是描述推论的主要工作,因为均值和方差也是随机变量的系统性特征(见 2.2 节)。其次,它使得我们能够把一个因果推论问题分成系统与非系统两部分。研究者可能对随机变量的系统性特征更感兴趣,在上述例子中就是单位 i 中的平均因果效应。为了让读者理解以上表述,让我们重新回到纽约选举这个例子。

随机变量是指假设同一竞选可以重复多次,民主党候选人(现任国会议员与否)获得的选票比例。对于非现任候选人,随机变量的期望值或者说该候选人所得选票比例的期望值为:

$$E(Y_4^N)=\mu_4^N$$

相应地,现任候选人得票比例的期望值为:

$$E(Y_4^I)=\mu_4^I$$

候选人是否现任在选区 i 中的平均因果效应就是随机因果效应中的系统性特征,我们将该效应定义为上面两个期望值的差(为表述的一般化,同样用单位 i 代替第四选区):

$$\begin{aligned}\text{在单位 } i \text{ 中的平均因果效应} &\equiv \beta \\ &= E\,(\text{单位 } i \text{ 中的随机因果效应}) \\ &= E(Y_i^I - Y_i^N) \\ &= E(Y_i^I) - E(Y_i^N) \\ &= \mu_i^I - \mu_i^N\end{aligned} \tag{3.3}$$

公式的第一行 β 即平均因果效应。在第二行中,我们将单位 i 的平均因果效应定义为随机因果效应的均值(也就是期望值)。第三行和第四行则告诉读者如何计算这个平均值。最后一行是平均因果效应的另一种表述方式(两个随机变量差的平均值等于各自平均值的差)。综上所述:当解释变量被赋予两个不同的值时,因果效应就是这些值对应的观

察值中系统部分间的差异。

公式(3.3)的最后一行看上去与公式(3.1)很相似。读者需要注意的是，该公式依然解决不了上文提到的因果推论的根本问题。事实上，用公式(3.3)的表述也许会使问题变得更复杂：即使可以通过“实现了的因果效应”来规避该根本问题，推论中仍然有很多一般性问题无法被解决，包括如何区分随机因果效应中的系统和非系统部分。在这里，我们采用霍兰的“因果推论的根本问题”这一说法，除了用来指代他提出的根本性问题之外，也包括推论过程中我们提到的其他问题。3.3.2 小节将用更为一般化的公式来表述因果效应，该公式化表述在本书的各个章节将被经常提到。

随机因果效应中的其他系统性特征也可能引起我们的研究兴趣。例如，我们用 2.6 节的公式(2.3)来表示获得选票的方差，如果研究者希望进一步了解民主党候选人的现任身份在选区 i 中可能的(或实现了的)因果效应的方差，便可以通过如下表达式获得：

$$\text{单位 } i \text{ 中的因果效应的方差} = V(Y_i^I - Y_i^N)$$

在计算因果效应方差 $V(Y_i^I - Y_i^N)$ 时，上述公式表述可以避免在计算过程中引进新符号。新的现任议员当然很希望了解现任与否对选票比重作用的方差，这样就可以因此判断他们未来的经历与前任议员间的相似度，同时也能据此了解自己的任何决策判断需要在多大程度上依靠以往的效果。研究者必须明白因果效应的方差虽然可以通过估计获得，但它本身作为现实世界的一种存在因而是确定的，记住这一点非常重要。

3.1.2 一个定性例子

我们在 3.1 节中虽然已经给因果关系下了明确的定义，但涉及的某些概念依然有些抽象和复杂。为了便于读者进一步理解，我们在本小节用一个简单的定性例子，并从定量角度对因果关系作进一步的阐释。通过这个例子，读者可以在理解我上小节提到概念的同时，不必拘泥于定性研究所推崇的语意细节和文化敏锐度。

除了我们关注的那个解释变量之外，如果历史能够在其他因素不变的前提下重演，研究者就能找到很多问题的答案。政治学家关注的一个重要问题是某项法律的实施对政治及政府的影响。例如，国会通过一项税法，其目的是通过增加税收或者改变居民消费方式来吸引投资。该项法律能否获得预期效果呢？研究者及政策制定者可以观察税法施行之后是否达到预期的效果；但即使我们确实观察到了预期的效果，我们能保证说这一定是由于新税法的缘故吗？未必，如果国会在同时期也调整了投资政策的话。因此，一个最有说服力的设定是让历史重来，依次观察实施与不实施这项税法后的不同结果，两者间的差别就是新税法的效果。在现实世界这当然这是无法实现的，但其蕴含的逻辑却有助于研究者的研究设计并由此获得一个近似答案。

让我们从比较政治学的角度讨论下面的例子。随着苏联的解体，东欧以及苏联共和国的政府开始着手构建新的政治体制。这些国家进行了一系列他们自认为很伟大的政治尝试：颁布了新宪法以期建立稳定的民主体制。其中涉及一个宪政层面的问题：到底选择议会制还是总统制。即使在学界，关于哪种体制能带来稳定的民主的争论也非常激烈且没有一致结论（Linz，1993；Horowitz，1993；Lijphart，1993）。当然，具体论战的内容非常复杂：议会制与总统制下面还能继续细分出不同类型，同时这两个制度也和其他制度相关，比如选举制度的类型。本书的目的并不是对所有相关选择做出全面的分析，我们试图简要地讨论不同的制度，重点是让读者可以通过这个定性例子定义出因果效应。为了实现这一点，我们需要区分出因果效应中的系统特征与非系统特征。

总统制和议会制孰优孰劣的争论涉及这两种制度的不同特点。在以下讨论中，我们主要将注意力放在两点上：第一，哪种制度能最广泛地代表选民利益；第二，哪种制度可以产生有决策力的领导人。就第一点来说，在议会制下，各种群体都能选出代表在议会中占有一定数量的合法席位，因此该制度能更好地代表社会各个群体并反映出他们的利益。相比之下，总统制要么全赢要么全输的特点必然会使得一些团体感到自己被排斥在政府之外，由此而生的疏离感会引发

不稳定；当考虑到第二点时，议会制恰恰是由于它要兼顾社会各种群体的利益，其决策效率和影响力会受到损害，这同样会导致不满情绪并影响社会稳定。①

同样，我们用这个例子来演示因果效应的定义。首先，设想建立一个议会制国家并在接下来的十年内定期衡量政体稳定性。我们通常用“民主政治得以持续还是失败了”、“未遂政变数量”等指标来衡量政体稳定性。与此同时，在同一个国家的同一个时期再选择总统制，也用同样的方法在同期内衡量它的稳定性。那么总统制下稳定性和议会制下稳定性的差异就是“实现了的因果效应”。读者已经知道这种方法无法实现，又一次凸显出因果推论中的根本问题。

此外，研究者也需要在这个因果关系中区分出系统和非系统部分。如果试验能被重复多次，那么平均因果效应就是由每次实验得到的“实现了的因果效应”的平均值。通过这个过程，研究者可以剔除那些非系统性特征，使平均因果效应只保留系统性特征。需要指出的是，系统性特征包括议会制下决策的优柔寡断或总统制下大多数人的不满意。而非系统性特征可能是总统的突然发病造成的政府混乱。显而易见，总统意外发病当然不是总统制的惯有特征，它可能偶然出现在某次试验中，但不会每次都有。②

这个例子另外一个让人感兴趣的方面是因果效应的方差。任何一个政府在选择政治体制时都会考虑体制是否能维护民主政治的稳定。但在实际情况中，一个国家只有一次选择，也只能进行一次尝试。在这种情况下，政治领袖关注的并不仅是平均的因果效应，也许更希望了解这个效应的最大值、最小值以及方差。一般认为，总统制会降低民主政治的稳定性，同时该作用的变化幅度很大：有时会显著地加强稳定性，

① 当然，这些差别本身也存在争议。一些人认为总统制也可以代表不同利益，也有人认为议会制在决策效率上同样能表现出色。

② 系统性特征和非系统性特征之间并没有严格的界限。比如，在总统制下，总统突然发病就是非系统性特征。但另一方面，总是受领导人难以预料的健康状况及其人格因素影响的确是总统制的弱点，这就是一个系统性特征了。同时，该特征会增加一些非系统特征出现的概率。

有时则从根本上破坏了稳定性。我们将这种不确定性称为“政治风险”。在这种情形下,民众和政治领袖也许更倾向于这样的方案:平均来说,该方案促进稳定的作用也许不太大,但作用的方差也相对小。这样的优势显而易见:能最大程度地降低社会动荡。

3.2 因果关系的其他定义

在 3.1 节中,我们给因果效应下了这样的定义:当因果变量有两个不同的值时,平均因果效应就是不同值对应被解释变量中系统部分间的差异。本节的任务就是通过该定义来廓清一些相关概念和观点。这些观点有些已被其他学者提及,它们包括“因果机制”(causal mechanisms)(3.2.1 小节)、“‘多重’因果关系”(“multiple” causality)(3.2.2 小节)、“对称”与“非对称”因果关系(symmetric versus asymmetric causality)(3.2.3 小节)。需要指出的是,这些概念和本书对因果关系的界定并不冲突。

3.2.1 “因果机制”

许多学者认为因果关系的核心是“因果机制”(见 Little, 1991: 15)。这种观点非常容易理解:任何合乎逻辑的因果关系都需要指出其作用是如何产生的。例如,研究者试图了解一个新的双边税务条约对降低美国对日本的经常项目赤字的影响。根据定义,该因果效应是指与相同条件下(同一国家在同一时期内)不实行该条约相比,实行该条约对经常项目赤字的影响。那么,对于该因果机制的解释就可能包含税约的制定及审批、针对条约的新闻报道、访谈跨国公司的相关人物、对减少涉及公司国际税负总量的补偿性举措(比如在两国之间改变转让定价规则或转移制造工厂)、那些受益于两国间资本和人力流动的其他公司及工人的进一步行动等。研究者通过对上述不同方面的分析,

进而了解该条约影响美日两国间收支平衡的具体机制。

从因果关系的实现过程来说，对因果机制的强调是有其合理性的：任何合乎逻辑的因果关系都需要指出其作用是如何产生的。识别因果机制是实证分析中很普遍的方法。该方法有时也被称作“过程追踪”(process tracing)，“历史分析”(historical analysis)以及“详细案例研究”(detailed case studies)(本书6.3.3小节将详尽地讨论这些概念)。许多优秀的案例研究都会涉及对因果机制的识别。

当然，对因果机制的识别离不开因果推论，在下面我们将对此展开讨论。研究者应该首先定义并估计出所有潜在的因果效应，然后在此基础上讨论每个因果效应对应的因果机制。如果要描述那些内在一致的因果机制，对于其中涉及的每个因果链的描述都需要借助本书在3.1节中对因果关系的基本定义。

因此，对因果关系的定义应该在逻辑上先于对因果机制的识别。在社会科学中，任意因果链间都会存在无穷多的因果步骤。如果我们认定一个解释变量会影响被解释变量，研究者就需要描述并识别出这两个变量间的所有因果联系。整个因果链条中涉及的相关变量间有很多组合，虽然有些组合并不是假设的因果关系，研究者也应当对此进行识别。这种方法意味着研究者需要进行非常多的回归，但回归的结果并不能定义出一个清晰的因果关系。

再回到之前政治制度和稳定性关系的例子(3.1.2小节)，我们先假设因果机制包括总统制下少数派别的不满意和议会制下的低效率。这些假设对政治稳定的影响是可以被直接观察到的。比如通过对比少数派的态度和政府决断力在两种制度下的差异。这些差异意味着两种制度对政治稳定性的具体影响机制是不同的，但这并不影响两种制度和政治稳定之间的因果效应。因此，对因果效应的定义并不需要我们一定理解涉及其中的所有相关因果机制，但如果没有前者，便不可能识别出因果机制。

用通过原因找结果的方法识别出因果机制并用来支持研究者想建构的理论，在我们看来是很有帮助的尝试。将不同层次的观察值都作为可以检验理论的现象，由此识别出的因果机制可以增加理论的解释

力。对因果机制的识别过程也可以启发研究者,使其发现新的因果假设。尽管因果机制的识别过程非常有用,但研究者不能将它和因果关系混淆。

3.2.2 “多重因果关系”

查尔斯·拉金(Charles Ragin)在他的著作(1987:34—52)中讨论了在较多解释变量和较少观察值的情况下进行研究的方法,以论证他提出的“多重因果关系”概念。他认为“研究对象也许有众多决定因素”,这和穆勒(Mill, 1843)提出的“原因的多样化”(plurality of causes)相似。该论点也涉及“通用系统理论”(general system theory)(George, 1982:11)中的“等效”(equifinality)原则。在存在多重因果关系时,这些学者都认为结果可以由不同的解释变量综合导致。①

在实际情况中,确实存在不同解释变量都影响同一个被解释变量的情况,但有些统计方法却错误地认为其中一些解释变量没有作用。拉金在他的书中就提醒研究者注意这个问题,在这点上无疑他是正确的。拉金指出,一些统计模型或定性研究设计并没有提醒研究者注意“多重因果关系”的存在。其实如果统计模型提醒的话,这个问题并不难处理。拉金的著作中就讨论了一些处理方法。

其实本小节讨论的“多重因果关系”的基本特征和我们对因果关系的定义是相容的。这两个概念不管是在定量研究还是在定性研究方面都差别不大。例如,假设一个人的收入水平同时被他自己和他的父母的教育程度所决定。只考虑其中一个条件显然是不够的。我们需要将所有可能的组合纳入考察范围:一类人本身受过良好的教育,同时他们

---

① 这点经常会被解释如下:没有解释变量被认为是解释变量某个取值的充分或者必要条件。对于这种表述,有些读者可能感到迷惑,因为当我们允许原因存在一定的不确定性时,必要条件和充分条件之间的区别就消失了。正如利特尔(Little, 1991:27)提到的:“让我们考虑这样一个说法:危机时期超级大国间的缺乏沟通会增加战争爆发的可能性。这是一个概率性的命题:解释变量(沟通缺乏)会增加结果(也就是战争)出现的可能性。但这种说法不能被理解成:缺乏沟通是战争的充分或必要条件。这种概率性的说法是无法被简化的。”

的父母也有较高的教育水平，另外两类人只具备其中一个条件，最后一类人两个条件都不具备。由于存在四种可能的情况，能否识别出“多重因果关系”就非常依赖于数据，但这依然不会改变因果效应的定义。根据定义，研究者还需要被解释变量：上述四种类型中每一个人的预期收入。

如果不同的解释变量产生了相同的因变量，为什么会出现这样的结果呢？比如，我们用“是否有大学文凭”（解释变量）将一个工厂里面的所有工人分成两组：有大学文凭的人和没有大学文凭的人。虽然在文凭上存在差异，但这两群人也许在工资（被解释变量）上并不会有什么差别。一个可能的原因是研究者考察的大学文凭对工人收入没有任何影响：是否有大学文凭对工人在工厂中的表现没有帮助。此外，也有其他因素能导致不同学历的工人拥有相同收入。大学毕业生由于教育程度更高而能获得额外收入，但那些未受过大学教育的人却拥有额外四年的工作经验。显而易见，工作经验有助于提高收入。在这种情况下，对于那些希望进入工厂的大学生们，研究者能告诉他们“大学文凭”对收入水平没作用吗？

因果关系的定义要求研究者谨慎地考虑反事实样本。在这个例子中，关键解释变量包括两组人：(1)有大学文凭的工人；(2)没有大学文凭但有四年工作经验的工人。被解释变量是工人的起薪。相应地，对该因果效应的定义就是：先记录大学毕业生的薪水，然后回到四年前让此人不上大学而直接进工厂工作四年后记录他的收入，这两个收入间的差距就是平均因果效应。在上文中，我们认为该效应为0，但这并不意味着大学教育对收入没有影响，除非(1)和(2)对应工人的平均收入差别为0。实际上，“大学教育对收入的作用”在逻辑上并没有唯一的定义，因为研究者识别出因果关系需要至少两个条件下的样本。当然并不是所有的研究问题都需要这两个条件，但不管怎样，它们都需要被清楚地识别出来。

另外一个可以获得因果效应的设定是比较那些有大学文凭的工人和那些没有大学文凭但也没有工作资历的工人。从某种意义上说，这是不可行的：那些没上大学的人在这四年中会有一些可以增加工作资

历的经历。但我们可以不让这个人从事一些与工厂工作无关的职业。和之前设定不同的是，该设定中，对于那些有大学文凭的样本的反事实是工作资历相同但没有大学文凭的工人。如果这两组人的工作资历不同，那么该研究设定获得的因果效应就和上段讨论的设定获得的结果相同。如果研究者本想控制工作资历却没有控制，基于此的实证分析将会由于“遗漏变量问题”而出现偏差。对这种类型的偏差我们将在 5.2 节中讨论。

虽然“多重因果关系”的识别对研究者的要求较高，但它的存在并没有改变因果关系的定义。事实上，在社会科学中，那些研究者感兴趣的被解释变量一般都受很多因素影响，即使这样，本书定义的因果关系依然成立。唯一需要读者注意的是，理解这种类型的因果关系需要给每个相关的因果效应精确地定义出反事实条件。本书第 5 章的内容将讨论如何预测因果效应，我们在那里会告诉读者，即使你们可以做到也不必对所有影响因素展开讨论。我们应当先集中关注所要研究的那个效应，在得到结论后再去讨论其他因素的影响(参见本书 5.2 和 5.3 节)。[1]

3.2.3 “对称”与“非对称”的因果关系

斯坦利·利伯森(Stanley Lieberson，1985:63—64)对他所提到的“对称”与“非对称”的因果关系进行了区分。他对解释变量增减相同幅度却导致了不同的因果效应非常感兴趣。用他的话来说：

---

① 对观察值中系统和非系统两个组成部分的强调，很大程度上源于我们的世界观：世界是或然的而不是必然的。因此我们不同意拉金的假设：“因为一个现象中的每个事件都是可被检验和解释的，因此通过比较方法获得的解释并不都是或然的。”(Ragin，1987：15)尽管研究者能够收集到一个现象中每个事件的信息及解释变量取值的所有排列组合，但这些信息和数据仍然由一些或然过程(见本书 2.6 节对这些过程的定义)所产生。这正是本书所要强调的关于科学推论的逻辑，基于数据的研究同样需要这一逻辑；该逻辑也许会使拉金的“布尔代数法则”(Boolean Algebra)失效，该法则是理论设计及因果推论的一般性方法。尽管如此，拉金所提供的方法作为形式理论的一种仍然具有价值(见本书 3.5.2 小节)：它使得研究者能够详细地界定出理论及其含义。

> 在检验 X_1(解释变量)对 Y(被解释变量)的因果效应时,我们也应当考虑当 X_1 增减相同的数值时,Y 值的相应改变是否相同……如果 X_1 和 Y 的因果关系是对称的或是完全可逆的,那么如果 X_1 变回原值(假设其他条件都保持不变),由之前 X_1 变化所导致的 Y 变化就会消失。

为了理解上述观点,让我们看这个例子:假设在 1998 年纽约第四国会选区的竞选中没有现任候选人参选,民主党候选人获得了 55%的选票。如果获胜的民主党人仍然参加下届(2000 年)选举,记住那个时候他的身份是现任议员了,根据利伯森的理论,现任与否的因果效应应当是在 2000 年选举中因为现任身份获得的额外选票份额。再下一轮(2002 年)的选举结果可以用来判断该效应是不是"对称的":如果 2000 年选举民主党因为没有当选而以非现任议员的身份角逐 2003 年的选战,那么候选人的得票率应当依然是 55%;反之,如果现任民主党人因为筹集了更多的竞选资金,改善了竞选组织同时没有非现任候选人参选,该党候选人也许能获得超过 55%的选票,这种情况下该效应就是"非对称"的。

尽管利伯森的论证过程非常清晰,但在我们看来,他的论证并没有构建出因果关系的定义,仅仅涉及因果推论:从已有的信息中获得因果效应的过程。我们在 3.1 节中给出了任一单位上的因果关系。就这个例子来说,基于 1998 年纽约第四国会选区的竞选,按照这个定义出的因果效应是:在相同竞选活动、相同时间及相同地区的条件下,民主党是不是现任国会议员在选举中获得的选票比重中系统部分间的差别。

但利伯森的表述并没有对假设进行限制,因此并不能定义出因果。这个例子中涉及的假设条件是:较之另外两次选举中由现任转变为非现任,如果解释变量由非现任转变为现任,将会出现什么情况?对这个例子的任何实证分析工作将涉及推论的很多问题,本书第 4 至 6 章会集中讨论这些问题。在当前这个例子中,读者也许会问是否由于第四选区中新注册了大量选民,从而导致估计效果更大(高估)?或者是由于我们忽略了在第一轮选举中民主党落败的那些选区,从而导致现任民主党候选人支持率低于应得支持率(低估)?

总之，在因果推论的过程中需要考虑利伯森的“对称”和“非对称”观点，但这些概念不应与3.1节中如何定义因果关系混为一谈。

3.3 估计因果效应需要的假设

研究者如何才能避免因果推论的根本问题，同时解决推论过程中系统性和非系统性部分的划分问题？本书第4至6章的内容将给出全面而详细的回答，在这里我们给读者做一些概括。处理好推论的根本问题需要两个假设：单位的同质性（unit homogeneity）（3.3.1小节）和有条件独立（conditional independence）（3.3.2小节）。和任何其他试图避免因果推论根本问题的尝试一样，这两个假设通常是无法检验的。因此，研究者有责任将研究设计中那些不确定的地方告诉读者。因果推论的工作并不是变戏法。对于这两个假设的详细讨论和清晰论证都应当基于研究者掌握的资料及现有研究。

3.3.1 单位同质性

通过以上的论述，读者应当知道解决因果推论根本问题的方法是历史可以在同一时间和同一地点重现，这样研究者就能给解释变量赋予不同数值，但这在现实世界无法实现。对于研究者来说一个次优的方法是：能在两个“同一”的单位上重复试验。当解释变量取一个特定值，如果与这两个单位相对应的被解释变量的预期值也相等的话，我们就称这两个单位是同质的（即 $\mu_1^N=\mu_2^N$ 和 $\mu_1^I=\mu_2^I$）。举个例子，我们观察到在第一选区的候选人情况是 $X=1$（现任候选人）而第二选区是 $X=0$（非现任候选人）。如果单位是同质的，那么就能直接用这两个选区的选票比重计算出因果效应 β。在候选人情况相同的条件下，β 在这两个选区中也应该是相等的。如果样本包含 n 个观察值，单位同质性的意思就是那些取值相同的解释变量所对应的被解释变量的期望值

也应相同。当然,这仅仅是一个假设:这两个选区在某些未知方面也许有诸多的不同。因此,如果假设和现实不符,就会导致因果推论产生偏差。事实上,任何现实中的两个选区都有或多或少的差异。如果选举活动在不同选区重复进行,同质性假设就要求这些选区在各个方面都要相同。例如,如果下雨会减少选民人数,同质性假设就要求下雨造成的影响不会因为选区的不同而存在差异,除非两个选区的气候特征存在系统性差异。

在下面这段引文中,霍兰(1986:947)用一个例子来说明单位同质性假设。需要提醒读者注意的是,霍兰是从"实现了的因果效应"而非"平均因果效应"的角度展开论述。由于例子中提到的实验几乎不存在随机性,所以霍兰的单位同质性定义非常接近我们的定义。(事实上,单位数量越少,随机的可能就越低,此时同质性假设就越发有用。这点我们将在4.2节中讨论)。

> 如果"研究单位"是房屋里面的一个房间,t(介入变量)表示按下这个房间电灯开关,c(控制变量)表示没有按开关,"被解释变量"则表示在 t 或者 c 发生之后电灯到底是亮还是灭。我们按下开关就可以知道 t 和 c 对应的被解释变量的取值了。显而易见,这样的假设是比较可靠的,大部分人都有这样的常识。但正当我要着手实验时,电灯却无来由地亮了或者熄灭了。那么知道 t 和 c 就一定知道灯泡是否亮就值得怀疑了,除非我们再想出一个新的实验方案做到这一点。

在这个例子中,同质性的假设是如果两次实验中都按下开关(霍兰用 t 表示)所获得的预期值(电灯是否会亮)应该是一样的。相似地,同质性也假设如果两次实验中开关都是开着的(用 c 表示),灯泡是否亮的预期值也将是一样的。注意在第一次实验后需要将开关重新调到关闭状态。但我们也要留意这样的情形:在第二次实验中打开同一个开关对灯泡是否亮没有任何影响(比如第一次实验后保险丝就熔断了)。一般而言,如果研究者只有一个样本,那么单位同质性的假设就无法得到检验(尽管如此,研究者可以检查墙内的线路进而提出一系列关于因果机制的新假设)。

单位同质性是一个比较强的假设,也许并不符合现实情况。如果我们将这个假设稍微放宽一下,得到的版本叫做“同效应假设”(constant effect)。单位同质性假设相同的解释变量在不同的观察值中对应的被解释变量的预期值需要相等,而同效应假设仅要求因果效应相同。也就是说,即使两个观察值对应的被解释变量不同,只要他们之间的差别相同就满足该假设了。因此,解释变量在同方向变化相同幅度,如果其对应的被解释变量的预期值也变化相同幅度,这就符合同效用假设但并不满足单位同质性假设。这也是为什么我们认为不变效应假设没有单位同质假设那么严格。举个例子,民主党非现任候选人在两个选区中获得了不同的选票比重(比如 45%对 65%)。只要现任议员的身份在两个选区中都能增加 10%的选票比重,那么我们就认为该假设成立。

不管是单位同质性假设还是同效应假设,都是科学研究的基础。它们都借鉴了比较案例研究的方法:比较不同解释变量值的样本并观察其相应被解释变量的取值。如果认为被解释变量的不同取值就是由于解释变量的不同所导致,这就意味着因果效应的存在。读者需要注意的是,在这种情况下,“认为”所依赖的正是单位同质性或同效应假设。

研究者完全可以突破时空的限制去寻找同质的研究对象。比如,可以对比同一个选区不同时期或者不同选区同一时期(或是这两种方式的其他组合)中民主党是不是现任议员对选票的影响。同质性假设通常是无法验证的,这也意味着因果效应只能被估计出来(estimated)而不能被确切地认定(know)。但研究者必须明确界定出假设的适用范围:“是不是现任医院对得票数的影响相同”在多大的范围内成立?在历次选举中吗,还是除了参议院选举以外的所有议会竞选?抑或仅适用于北部选区?还是仅适用于过去 20 年中的竞选?

此外,读者应该注意单位同质性假设与本书 1.1.3 小节中讨论的“复杂性”和“唯一性”之间的联系。我们在该小节中论证了社会科学研究的一般性依赖于我们简化事实的能力。将简化事实作为因果推论的目的意味着研究样本需要符合同质性假设。单位同质仅仅作为假设存在,现实世界很难达到该假设的要求;研究者无法将国会选举类比为电

灯开关,更不要说更为复杂的革命运动了。但读者需要知道的是,对研究样本同质程度的把握有助于我们估计推论中不确定性及偏差的程度。

3.3.2 有条件独立

有条件独立假设解释变量独立于被解释变量。该假设在概率论中经常会被用到,有时也出现在统计学中,不过含义不大相同。该假设的具体含义是在考虑(或控制)解释变量之后,解释变量的赋值过程独立于被解释变量(通常是两个,当然也可能更多):Y_i^N 和 Y_i^I。在这里,我们用"赋值"这个术语来描述变量获得特定值的过程。对于采用实验方法的研究者来说,给解释变量赋值意味着有些研究对象被纳入介入组,另一些被纳入控制组。在非实验环境中,解释变量可以被自然或环境"赋值"。但是这种赋值过程的关键在于解释变量的取值不能由被解释变量决定。当(至少部分的)解释变量的取值由被解释变量导致时,就会出现所谓的"内生性"的问题。对该问题的讨论见本书 5.4 节。

基于随机选择和赋值的大样本分析是满足有条件独立且不需要单位同质性假设的最可靠方法。随机选择和赋值的过程之所以有助于因果推论,是因为该过程本身满足三个假设。这三个假设构成了有条件独立的概念基础:(1)解释变量的赋值过程独立于被解释变量(也就是说不会有内生性问题);(2)不会存在 4.3 节讨论的样本选择偏误问题;(3)同时,在 5.2 节将要讨论的遗漏变量偏误也被克服。其实,只要能够满足以上三个条件,不论是通过随机选择和赋值(将在 4.2 节中讨论)还是其他方法,都可以避免因果推论的根本问题。

当然,有条件独立并不一定需要随机选择和赋值。即使解释变量的"赋值"过程不独立于被解释变量,只要研究者了解赋值过程并引入其他控制变量,有条件独立性依然能够被满足。让我们举个例子,如果想要估计约旦河西岸的居住隔离政策对当地巴以冲突的影响,仅从这两个变量的相关关系中去寻找因果效应将严重违背有条件独立的假设。因为那些选择居住在隔离区的以色列人和巴勒斯坦人也许是出于

信仰：只有自己才是西岸的真正主人。对双方而言，这种信仰的极端表现就是冲突的频繁发生。研究者无法区分到底是隔离政策还是意识形态的作用。区分这两个因素具有完全不同的政策性含义：如果隔离政策是导致冲突的原因，出台新的居住政策也许会缓解冲突；如果意识形态才是背后真正的驱动力，新政策不但不会获得预期效果，反而会进一步恶化事态。对于研究者来说，精确地测量出居民意识形态并加以控制可以处理这个问题：极端主义政党在以色列人中受欢迎的程度或巴勒斯坦解放组织在巴勒斯坦人中的受欢迎程度。当做完这一步之后，通过比较意识形态相似但处于不同隔离程度的社区，研究者便可以控制住意识形态对该因果关系的干扰。

如果随机选择和赋值不可行，那么研究者就无法控制分配和赋值的过程。在这种情况下，研究者就需要依赖于单位同质性假设来进行因果推论了。由于该假设在现实中很难达到，对因果推论过程中不确定程度的讨论便显得尤为必要。值得读者注意的是，单位同质性假设在本书 5.6 节讨论的对观察值“进行匹配”时尤为重要。

定义因果效应的公式化表述

为了便于理解本书后面的内容，我们现在给出定义因果效应的公式化表述。一般说来，一个随机变量 Y_i 通常会对应 n 种情况。在上面谈到的定量例子中，n 代表国会选区的数量(435)，y_i 是随机变量 Y_i 在选区 i 中的情况：民主党获得的选票比重(假设是 56%)。如果可以在选区 i 中不断重复选举，该区非现任民主党候选人获得的平均选票比重的期望值是 μ_i^N。我们将解释变量定义为 X_i，当选区 i 的民主党候选人不是现任员时，该变量取值为 0，反之为 1。此时单位 i 的平均因果效应为：

$$\beta = E(Y_i \mid X_i = 1) - E(Y_i \mid X_i = 0) = \mu_i^I - \mu_i^N \tag{3.4}$$

将该公式进行简单化整理可得：

$$E(Y_i) = \mu_i^N + X_i(\mu_i^I - \mu_i^N) = \mu_i^N + N_i\beta \tag{3.5}$$

此时，当选区 i 的民主党候选人不是现任议员时，$X_i=0$，将该值带入公式(3.5)中，可得期望值为：

$$E(Y_i \mid X_i=0)=\mu_i^N+(0)\beta=\mu_i^N$$

同样，当选区 i 的民主党候选人是现任议员时，期望值是 μ_i^I：

$$E(Y_i \mid X_i=1)=\mu_i^N+(1)\beta=\mu_i^N+\beta=\mu_i^N+(\mu_i^I-\mu_i^N)=\mu_i^I$$

这样，公式(3.5)就是一个关于因果推论的模型，其中的 β 就是因果效应，它表示两个理论上可以获得的得票比重间的差异。为了获得进一步推论，将公式(3.5)进一步简化。假设 Y_i 的均值为 0(当然也可以使用对平均值的偏离表示，这样做并不会限制模型的适用范围)，相应地，公式(3.5)可以简化为：

$$E(Y_i)=X_i\beta \tag{3.6}$$

参数 β 依然表示平均因果效应的理论值。作为随机变量的系统特征，该参数是因果推论的目标。以上公式表示是"回归分析"的一个特殊例子。虽然"回归分析"在定量研究中非常普遍，但读者需要注意的是，该方法获得的估计系数仅在某些时候才与真实的因果效应相同。

3.4 因果推论的判断标准

在以上章节中，我们提到通过随机变量定义出因果关系的过程可以使读者严格地区分该过程和其他系统性特征，如平均值或方差间的区别。但研究者还是可以通过这些系统性特征进行描述性推论。在 2.7 节中，我们用无偏性和有效性来评价描述性推论，这些标准同样可以被用来判断因果推论。因此，第 2 章的绝大部分内容同样适用于对因果推论的讨论。在这里，我们将简单总结这两种推论间的差别。

2.7 节提到描述性推论的目的是获得平均值(也就是随机变量的期

望值)。该期望值用 μ 表示。μ 是一个固定的未知数字。如果 μ 的估计值与多次重复同一试验获得的 μ 均值相等,那么 μ 的估计值就是无偏的。

类似地,我们定义随机因果效应的期望值是 β,它也是一个固定的未知数字。和上面均值无偏性的定义类似,对它的无偏性定义为:如果 β 的估计值与多次重复同一试验以后获得的 β 平均值相等,那么 β 的估计值就是无偏的。

对有效性的定义也类似,该标准被定义成多次重复相同试验中的差异。以上概念都非常重要,在研究者处理因果推论中遇到的问题是,它们将提供依据。本书第 4 到 6 章将具体讨论这些内容。以下是对这些概念的公式化表述。

因果估计无偏性的公式化表述

以下将用公式表示从 3.1 节开始涉及的因果效应中参数估计值的无偏性。其中用到的符号和逻辑与 2.7 节中描述性推论无偏性的公式化表示相互对照。为了简单起见,我们只考虑含有一个解释变量和一个被解释变量的简单线性模型,表示如下①:

$$E(Y_i) = \beta X_i$$

用最小二乘估计法获得 β 的估计值:

$$b = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} \tag{3.7}$$

为了确定 b 是不是 β 的无偏估计值,需要对 b 求期望值——多次重复试验获得的估计值的平均值,该期望值是:

$$E(b) = E\left(\frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}\right) \tag{3.8}$$

① 为了避免使用常数项,我们假设所有变量的均值都为 0。该假设可是使表述变得简单,但并不会影响所获结论的适用范围。

$$= \frac{\sum_{i=1}^{n} X_i E(Y_i)}{\sum_{i=1}^{n} X_i^2}$$

$$= \frac{\sum_{i=1}^{n} X_i^2 \beta}{\sum_{i=1}^{n} X_i^2}$$

$$= \beta$$

该公式证明了 b 是 β 的一个无偏估计值。

估计有效性的公式化表述

本部分将评价 3.1 节中 β 估计值的有效性。公式(3.8)已经证明了该估计值是 β 无偏估计值，该值的方差是：

$$V(b) = V\left(\frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}\right) \tag{3.9}$$

$$= \frac{1}{(\sum_{i=1}^{n} X_i^2)^2} \sum_{i=1}^{n} X_i^2 V(Y_i)$$

$$= \frac{V(Y_i)}{\sum_{i=1}^{n} X_i^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$$

该估计值的公式表达式是两个部分的函数：第一，数据中每个单位的随机性越大（这意味着 σ^2 越大），b 的方差就越大。这可以从公式(3.9)的最后一行看出；第二，解释变量的方差越大，b 的方差就越小。在极端情况下，如果 X 没有任何变化，研究者就无法估计解释变量对被解释变量的作用。在这种情况下，b 无限大，即作用完全不能确定。从这些讨论中读者应该知道，选择变化范围较大的解释变量能够获得更高的推论有效性。在实际操作中，研究者可以采用这个标准去评估不同情况下获得的因果推论有效性的高低。用一个很生动的例子来说明这一

点:用一把尺子在纸上画一条直线使其通过两个点。如果这两个点非常接近(这意味着 X 的方差很小),没画准的可能要大于两点距离很远时的情况(这表示 X 的方差很大)。

3.5 建构因果理论的规则

很多针对定性研究的建议都极其详尽且富有建设性,但这些建议一般只涉及研究的某些特定方面。在这些问题的解决过程中,研究者心里一定要有全局意识:针对那些特定问题的解决方案必须有助于整体上的因果推论。我们已经给出了因果效应的理论定义,也讨论了因果推论过程中涉及的一些问题。在这一小节中,我们将提供一些理论建构的一般化规则。即使研究者已经进入数据收集阶段,但针对理论的改进工作并不会停止(针对此点的讨论见本书 1.2 节)。

我们建立因果理论去揭示一个现象或者一系列现象发生的原因。任何理论,不论旨在归纳还是演绎,都包含一系列相互联系的因果假设。这些假设是为了详细说明那些能够产生可观测后果的变量间关系:如果解释变量取一个特定数值,就可以预测出其对应的被解释变量的值。检验或者评价任何因果假设的工作都需要因果推论。建立在假设之上的理论应该保持内部一致性(internally consistent),否则产生的假设可能会相互矛盾。

上述定义对研究者的理论及假设建构工作具有广泛的指导意义。为了更好地构建出理论,我们将在以下部分给出五个规则供读者参考。为了便于读者理解,每个规则均被辅以具体的例子。

3.5.1 规则 1:建构出能被证伪的理论

该规则并不是说一个没法被证明是错误的“理论”就不是理论。我

们的意思是研究者建构的理论需要被尽快及尽早被证明是错误的。这也并不意味着研究者要尽力去犯错。在我们眼中,即使一个错误的理论也比一个无法知道对错的理论要好。对于理论的可证伪性的强调促使读者要用正确的态度对待研究的不确定性:谨慎对待理论并防止将其教条化。如果有足够的科学证据来质疑最初的理论,研究者应该随时做好抛弃这些理论的准备。对待任何理论(或是从该理论获得的假设),研究者都应该自觉地询问:哪些证据可以证伪它?该问题适用于所有的理论及相关假设,但首先,回答这个问题的人应该是研究者自己。

卡尔·波珀深入地研究了这个问题(Popper, 1968)。波珀认为,证明一个理论成立(证实)和证明一个理论不成立(证伪)之间有着根本性的不对称关系:前者和科学几乎无关,后者却是科学的关键。按照波珀的观点,理论被建构出来后就成为科学知识的一部分而被接受。理论具有普遍意义,但涉及的假设却是特定的。从技术上来说,理论可以包含无穷多的假设。然而,在实证检验过程中,研究者只能检验其中有限的几个假设,这意味着"理论是不可能被证实的",研究者不可能检验一个理论所有能被观测到的结果(Popper, 1968:252)。无论多少个假设的检验结果和理论一致,对于一个已经被接受的理论来说都改变不大。但哪怕其中有一个假设被证明是错误的且与理论不一致,该理论就会被证伪并从科学知识中剔除。正如米勒(Miller)的表述:"通过了检验并不会给一个假设带来什么明显的改变,但如果没通过检验,即使只有一次,却有很大的差别。"(Miller, 1988:22)当然,波珀并没有说证伪是一个明确的概念。他承认任何基于实证的推论工作都有不同程度的不确定性(Popper, 1982)。在讨论"失验"(disconfirmation)这个概念时,他这样写道:"即使承认'证实和证伪之间'存在不对称,我们也会发现理论不大可能被明确地证伪。"(Popper, 1968:42)

波珀的上述观点对构建理论非常重要:我们应该构建那些容易被证伪的理论。此外,波珀对理论不确定性的强调也值得读者注意。尽管如此,在对现有社会科学理论进行评价时,我们就会发现证实和证伪间的非对称关系并不是非常显著。两者都被视作对现有科学知识的补充。一般来讲,问题的关键并非在于理论正确与否——事实上每个理

论被观察到的结果中总有或多或少的错误——而是该理论能在多大程度上解释世界。按照波珀的标准，基于理性选择假设的理论在很久以前就被证明是错误的，因为有很多事实和理论预测的不一致。尽管如此，社会科学家并没有抛弃理性选择假设而是对其进行修正，其原因就在于该理论对很多问题依然有解释力（参见 Cook & Levi，1990）。除了这个例子，对理论的这个判断标准也适用于其他社会科学理论。对理论进行证伪的过程其实是寻求理论适用限度的过程。其实，不管被观测的结果是否和理论预测一致，我们都能在这个过程中受益。

对于科学家，尤其是社会科学家来说，波珀的非对称观点对于评价公式化理论来说也许并不十分有用。比如，对波珀观点在物理学上的应用，奥黑尔（O'Hear，1989：43）认为：

> 波珀总是倾向于那些具有普遍意义的理论。但我们必须重申：提出并检验那些普遍性的理论仅仅是科学的目标之一。时间和空间差异会导致很多条件发生显著变化，因此，真正的普遍性理论也许根本就不存在。对此我们必须予以足够重视。尽管如此，科学依然给大家贡献知识并给我们周遭的世界做出各种预测。

我们认为，上述基于物理学的观点一样适用于社会科学。

此外，波珀关于理论的评价方法并没有区分刚构建的理论和已经被检验无数次的理论。对于区分普遍性理论到底是真理还是虚构时，波珀的理论是适用的；但如果着眼点是寻找理论的适用限度的话，他的观点就不那么有用了。在本书中我们多次提及，任何一个特定假设的推论只能通过陈述猜测（或估计）及测量该猜测的不确定性来获得。不管推论是否和理论一致，推论发现都会影响理论的建构和估计的确定性。①

---

① 有些人可能认为（也许是指责）我们是“辩护主义者”或者“机会主义的辩护主义者”（Lakatos，1970）。如果一定要给我们打个标签，我们倾向于更加有条理及哲学意味浓厚的贝叶斯（Bayesian）标签（见 Leamer，1978；Zellner，1971 以及 Barnett，1982）。事实上，我们的目标与波珀的观点存在差别。假设我们认同波珀的目标，我们一般也会认同他的推理过程；同样，如果他认同我们的目标，他估计也会认同我们的推理过程。尽管如此，我们还是相信我们的目标更加接近社会科学研究并且更易获得成功。

让我们看这个例子:假设在美国总统竞选中,不管是民主党还是共和党的竞选策略,都对竞选结果的影响不大。从该假设其实可以引申出一系列具体的假设,比如,电视和电台播出的竞选广告、两党提名候选人的竞选辩论对选民决定的影响都不大。对上述较为一般化的那个假设的检验其实都要转化为这些具体假设的检验工作。针对该问题的研究表明,从大选开始前举行的两党大会中获得信息可以非常准确地预测选举结果(Gelman & King, 1993)。这种检验工作可能和理论一致(毕竟如果在竞选前就能预测竞选结果,竞选本身对选举结果就没什么影响了),但也不是说一定要符合理论预期。大选时有很多因素能影响选举结果,并不是所有因素都会被理论考虑进来(正是这种忽略解释了预测有时会发生错误)。大部分研究者对这种预测的偏差都持有认同态度。当然,理论有时候准确地预测了结果,多数情况下,这纯粹是巧合或者决定那一年选举结果的所有因素都被研究者考虑到了。

研究者继续在预测模型中加入可能影响选举结果的因素,比如为竞选投放的电视电台广告时长、两党候选人的语言表达能力以及对辩论结果的判断等等。如果发现这些因素都不影响选举结果,按照波珀的观点,大选本身不影响选举结果的假设就依然成立。如果研究者进行大量相似的检验都没有证伪该假设,理论依然有错误的可能:毕竟我们都不可能穷尽影响选举结果的因素。同理,不管多少次检验和理论一致,总统大选会影响选举结果仍然有可能正确。

但是,如果竞选中的突发事件,比如媒体发现候选人的某种不道德行为,确实影响了选民的选择,上述理论就被证伪了。尽管此时理论还没有被确凿地证伪(在波珀看来这是不可能的),研究者从中学到的东西也多过那些重复多次但次次都和理论一致的检验。

但对本书的作者来说,这并不是社会科学研究应该采用的方式。如果一千次检验中有一次和理论预测不符,即使这次检验是有效且确定的,研究者也不要轻易放弃竞选活动没有作用这个假设。研究者要做的应该是进一步修正理论:除非忽然发现某一候选人不道德行为的大量证据,一般而言,大选本身对选民行为不存在任何影响。但读者应当明白,诸如此类的修正会限制理论的适用性,因此在做该修正之前需

要利用新数据对其评估。理论仍然具有解释力,同时对理论的检验也可以让研究者更多地了解理论的适用范围。针对理论的每一次检验不光要评估其有效性,也要评估其不确定性,只有做到这些,研究者才能了然其适用范围。

以上讨论对理论的建构非常重要,当然研究者应该了解运用上述方法时需要注意的事项。对于理论和假设中出现的例外,研究者也应当保持清醒的认识。以下我们将告诉读者如何扩大理论和假设的适用范围。虽然该策略十分有用,但我们必须牢记社会科学理论并不是放之四海皆准的。很多理论都声称具有普世性,比如马克思主义理论中的某些表述和理性选择理论,但这些理论不是采用同义反复的方式建构的(在这种情况下,既不能说它们正确也不能说它们错误),就是通过允许"失验"方式建构的(在这种情况下,理论无法做出正确的预测)。大多数社会科学理论只在特定情况下有效(比如在候选人没有不道德行为的选举中)或在一系列限制条件下有效(比如在发达工业化国家而不是一般工业化国家,在众议院而不是参议院竞选中)。在任何情况下,研究者都应当尽可能地去廓清理论或假设的适用界限。下一个步骤很自然就是为什么会存在这些界限?适应于参议员选举的研究发现为什么不适用于众议院竞选?为什么工业化水平可以影响因果关系?让研究结论更具有一般意义,研究者还需要考虑哪些变量?通过提出这样的问题,研究者就能不断超越理论或者假设的适用界限,进而寻找到其他可以扩大理论解释范围的因素。

有一点值得向读者说明,就是理论和假设的评估过程是很灵活的:某个特定的实证检验结果并不能立即肯定或者否定理论。当一个检验结果和理论预期不一致时,研究者不能马上就抛弃这个理论,而是应当想想出现这种结果是不是由于一些偶然性的原因。即使理论可能在某些特定环境中不适用,但通过一定的调整能保持理论的适用性。科学的进步正是通过这种调整得以实现。但读者应当意识到如果调整过度,理论就会变得无懈可击以至于无法被证伪。因此,研究者在保持理论和新证据一致时应当非常谨慎,最好不要增加太多的例外情况。

如果研究者只证伪了理论的某些方面,那么通过增加例外的方式

依然使得该理论成立。虽然这会减少理论的成立条件，但该方法依然可以接受。只是在这种情况下，理论的价值就降低了。按照本书的术语，我们认为理论的解释力度降低了。[①]读者需要注意，如果通过上述方法获得的“理论”只有无用的特例及例外的话，这个时候整个理论和假设都应当被抛弃。总之，如果理论包含太多例外，这样的理论便毫无价值。因此，简约原则，即理论中的组成越少越好，并不具有普遍适用性。研究者需要的不是简约原则，而是最大化理论的解释力度，只有在此基础上再只追求简约原则。在很大程度上，科学正是用较少理论解释较多现象的过程。该目标本身并不能辨别那些简约的理论是好是坏。最大化理论的解释力需要研究者尽可能地用较少事实去解释尽可能多的结果。当然，简约原则有助于该目标的达到，但不是在任何情况下都适用。在很多情况下，研究者为了更多地解释世界，需要复杂一些的理论。在这种情况下，当然应该采用非简约原则，因为它比简约理论更能最大化理论的解释力。[②]

3.5.2 规则2:建立内部统一的理论

一个内部不一致的理论不仅可能被证伪，理论本身甚至可能是错

① 和通常一样，当对某个理论作出调整以和搜集的证据一致时，研究者应该基于新搜集的数据评价该理论(或理论中与新证据相关的部分)。

② 对波珀观点的另外一种表述是“你不能证明一个否定的观点”。对此，他认为之所以不能这样做是因为结论和假说一致可能由于错误的检验工作所致。那些试图证明一个否定观点的学者常常遇到这个问题。该问题带来的麻烦不仅仅是理论上的，也会影响发表:杂志更愿意发表肯定的结论而不是否定的结论。

这就会导致所谓的“文件抽屉问题”(the file drawer problem)。对很多定量研究来说，该问题经常遇到。我们假设世界上不存在某个特定模式，那么对该模式的每100次检验中就有5次结果不在95%置信区间内，这就会产生错误的推论。如果我们假设学术期刊只发表那些肯定结论而非否定结论的文章，那么该期刊所发表的就是那些“显著的”5%文章。这也就是说，期刊只发表了那些得到错误结论的论文，文件抽屉里却塞满了结论正确的研究(见 Iyengar & Greenhouse, 1988,就该问题对统计文献所做的综述)。事实上，研究者都知道这种偏好，而且他们的行为也会受到影响。即使很多社会科学期刊的接收率正是5%左右，情况虽然不糟糕但也值得重视。我们认为如果读者能够接受如下观点，那么抽屉问题就可以得到解决:一个否定的结论和一个肯定的结论一样有用，两者提供的信息同样丰富。只要研究者提供关于不确定性的估计及测量，所获结论就都可以被接受。

误的。在这种情况下，我们并不需要借助任何实证分析就可以判断出理论的真实性：如果某个理论产生的假设彼此矛盾，那么在真实世界中就无法找到证据来支持这个理论。因此，我们应该保证理论的内部一致，但读者应该知道实现这个目标通常是很困难的。数学建模是实现内部一致性理论的方法之一。众所周知，公式化模型（formal modeling）的方法在经济学中已经发展得相当成熟，也逐渐被社会学、政治科学、人类学等领域所广泛接受（见 Ordeshook，1986）。在政治科学的理性选择、社会选择、空间选举模型、公众经济学和博弈论等领域中，学者们已经用这种方法建立了大量理论和假设并获得了很多重要的结论。简要地说，采用建模方法可以非常明晰地揭示出由文字陈述理论内部不一致性的地方。

尽管如此，和我们平时对假设的检验工作一样，公式化模型也需要实证证据来检验其预测。采用公式化模型的方式的确能够帮助研究者更清晰地推理，当然也有助于理论内部实现一致，但它本身并不能检验社会科学理论。我们需要知道，为了实现数学表达上的简化及实现均衡，社会科学研究中公式化模型都有一些假设。物理学家认为物理世界是一种数学结构，但不大会有人像物理学家那样相信政治世界也是一种数学结构。公式化模型仅仅是模型，模型就意味着它们往往比较抽象并和真实世界有所区别。在实际研究中，公式化模型必须依靠过于简化的假设进行预测，这意味着有些时候这些理论没有多少实证价值。和那些非公式化的社会科学理论相比，公式化理论更加抽象：它们不能对现实世界进行具体的预测。公式化理论成立所依赖的特定条件和现实条件即使有时较为接近，却并不完全一样。

在所有的研究中，将模型公式化的过程简化是非常必要的。但尽管如此，我们也应该对任何模型的预测抱有谨慎小心的态度。例如，在建模中，我们往往假设遗漏变量的问题对于模型的预测结果没有任何影响。该假设对模型的推导非常必要，在本书很多对定性研究的公式化过程中我们均会涉及此假设。但读者应当清楚这样的假设在真实世界并不成立，它仅仅是模型的特征。只有在这些遗漏变量和理论本身不相关时，该问题的存在才不会导致理论预期和真实世界之间存在差

别。在做模型推导时,研究者不必检验这些假设,但要知道在实证检验时对该假设的检验则非常重要的。当然,公式化模型是否有用并不取决于假设是否正确。但研究者不应该将那些还没有被检验或证实的假设用在实证研究设计中。与此相反,研究者应该尽量对公式化模型进行补充,使之可以引导实证研究。

好的公式化模型应当足够抽象,只有这样才能明确研究问题的关键特征,而对其进行推理时也较为容易。仍然以比例代表制度对政党影响为例,我们将在以下部分向读者展示如何对该问题进行模型化。首先必须明确该研究问题的关键解释变量是选举制度:比例代表制还是单一选区制。前者指的是政党获得议席的数量基于其得到的选票比重,而后者则每个选区只有一个获胜者。其次,该研究问题的被解释变量是政党数量,该变量一般用来测量政党分化程度。最后,该研究问题的假设是和单一选区制相比,比例代表制是否导致政党数量的增加。为简单起见,上述模型仅包含选举制度的一些必要特征及政党分化程度变量。因此,由该模型推导出来的结果仅仅是一个假设而非结论。在现实世界中,比例代表制对政党分化的真实影响应当通过定性或定量方法检验。

虽然该模型的用意在于探索比例代表制是否导致正当分化,同时模型中也没有引入其他影响政党分化的变量,但在对两者关系的实证检验中,如果仅仅使用上述两个变量则是不可取的。已有的研究发现在两者之间存在内生性问题(5.4 节):之所以按比例给党派分配议席,正是由于一国中的政党比较分化。在互为因果问题之外,遗漏变量造成的偏误也应该引起研究者的注意:一般来讲,那些在种族及宗教上较为分化的国家,其政党也分化严重,而这样的国家往往倾向于采用比例代表制。这样,遗漏变量之所以导致推论误差,是因为该变量与解释变量及被解释变量都相关。因此,如果研究者在实证检验中忽视了诸如种族和宗教因素,就会获得有偏误的推论。

综上所述,在描述研究思路及建立内部一致的理论时,公式化模型非常有用。对于那些比较复杂、由语言陈述的理论,只有使用公式化模型才能发现和改进内部存在的非一致性问题。此外,研究者需要明确的是实证检验并不能完全依赖公式化模型。因此,该方法并不能使我

们避免科学推断中存在的实证相关问题。

3.5.3 规则3:谨慎选择被解释变量

研究者在研究中应始终保持小心谨慎的态度。在选择变量，尤其是选择被解释变量时要尤其谨慎。就此，基于定量研究和定性研究中易犯的错误，我们提出如下三点建议。

首先，被解释变量应该是不独立的。研究者经常犯的一个错误就是选择一个能够导致解释变量变化的被解释变量。对于这种内生性问题导致的结果及规避该问题的方法，我们将在本书5.4节中讨论。在这里，我们仅仅提醒读者一点，就是尽量选择那些外生的解释变量和内生的被解释变量。

其次，不要选择那些会使被解释变量是常量的观察值。虽然大家都知道这种错误不该发生，但总是会有研究者选择那些根本没有变化的被解释变量作为观察结果(我们将在4.3.1小节中讨论这个问题)。就该点建议还有一个问题值得大家注意，就是虽然在研究设计中我们努力回避被解释变量是恒量，但在部分情况下它仍然可能变为恒量。例如，假设我们基于解释变量中的两类来选择观察值，结果发现被解释变量在这两种类型中没有任何变化。通过这种研究设计估计出的因果效应就是0。一般说来，只要我们不用某些预先设好了的特定标准，就不会出现这样的问题。

最后，被解释变量应当可以捕捉到研究者试图解释的变化。这条建议看上去十分简单，但在实际研究中完美地实现却有一定的困难，正如斯坦利·利伯森对此的阐述：

> 多伦多安大略科学中心(Ontario Science Centre in Toronto)主办的引力展览为大家提供了一个很有启发的例子。在展出中，一枚钱币和一根羽毛同时从一个真空管子中落下，它们差不多同时到达管子底部。由于管子并不完全是真空的，所以钱币比羽毛稍微早一点接触到管子底部。假设我们设计一个试验来研究不同物体的下落，但在试验中没有控制管子里是不是真空。我们知道

在管子中完全实现真空是很难的，这恰恰很像社会科学研究，在绝大多数情况下都难以实现完美的控制。如果研究者发现物体落地时间不同，他们就会去寻找导致落地时间不同的因素。例如物体的某种特性，比如密度或形状，这些因素在非真空状态下也许会影响下落速度。如果研究者足够幸运地考虑了所有因素并解释了不同物体下落速度的差别，那么他就有理由感到高兴并宣称所有影响物体下落速度的因素都被发现了。显而易见，如果研究者忽略了重力因素，那么上述实验设计就是有问题的(Lieberson, 1985:100)。

上面例子中的研究设定可能是错误的，除非研究者想考察的变量就是重力。如果重力是我们试图考察的解释变量但该变量在这个实验中却没有差异的话(因为该实验在同一个地点操作)，那么这个实验就不能检验出重力的作用。如果研究者感兴趣的是不同物体在不同气压下从相同高度落地所需的时间差异，那么利伯森描述的这个实验就非常重要。事实上，即使我们对重力非常了解，从这个实验中我们仍然可以获得重要的信息。不过，正如利伯森假设的那样，如果我们真的对重力造成的因果效应感兴趣，我们可能就需要一个因变量，它的观察值随着地球引力的变化而变化。在社会科学中，我们必须非常谨慎地保证我们关注的是研究的因变量，而不是研究设计中那些保持不变的背景因素。

对解释变量的无偏估计需要研究者将被解释变量的所有可能取值都纳入考察范围中。对被解释变量的取值范围所做的任何人为限制都可能存在选择性偏误问题(见本书 4.3 节)。例如，如果研究者试图研究冲突的爆发条件，那么我们就不能只选择冲突已经发生的样本作为观察值。基于这些观察值的研究虽然可以揭示出冲突差异的信息(正如之前重力试验中不同物体下落速度上的差异)，但并不能让我们了解冲突的爆发条件。如果你对爆发条件感兴趣，正确的研究设计应该是根据解释变量去选择样本并使其涵盖所有可能的情形：从没有冲突到有冲突。

3.5.4 规则 4：尽可能具体化

在之前的内容中我们强调理论可以被证伪、内部需要一致及被解

释变量要有变化，我们的最后一个建议是尽可能地将理论具体化。只要条件允许，我们都应当使用那些被观察的概念。但在社会科学理论中很多抽象的、难以在现实世界中观测到的概念却经常被使用，例如效用、文化、意图、动机、识别、智力等。虽然这些概念在建构理论的过程中非常有用，但除非研究者可以明确定义或测量出这些概念，否则研究者将检验采用上述概念的理论。因此，当使用诸如文化、国家利益、效用或动机来解释现象时，研究者在操作过程中应该谨慎并尽量独立于解释变量来测量上述现象。之所以保持谨慎是因为用这些术语来解释其他现象时，研究者不是在作同义反复就是没有区分出那些被观察到的现象。个人或国家行动可被解释成出于最大化效用目的去完成个人意愿或国家利益。但不管是最大化效用、实现意愿和国家利益，这些行为在事实上都要由个人或国家参与。因此，研究者基于此建构出的理论必须清楚且精确地阐述出哪些可被观察到的现象能证明其正确性并由此将其和其他理论区分开来。

当然我们的这点建议并不是说诸如意图和动机等概念就不重要了。这条意见是希望大家意识到衡量实证研究的标准是可以实证地证实或者证伪的。相较检验那些特定和具体的概念，为那些抽象的、无法观察和测量的概念提供实证证据非常困难也很难成功。归根结底，概念抽象及可被观察到的现象不清晰的理论很难被证伪。

研究者可以采用以下的折中策略：仍然从一个抽象概念开始。但需要意识到这个概念不能直接测量，需要做的就是用某些特定指标来替代它。和抽象概念不同，这些特定指标是可以被测量出来的。这样研究者就可以用这些特定指标从事解释工作。采用这种方法的唯一前提是那些替代指标一定能被观察到，例如在某种情况下可得的数据或者被记录下来的某次历史事件。这种折中策略在实证研究操作中有其必要性并被广泛采用。

尽管如此，这种折中方法也有不尽如人意的地方。比如，选取的指标和原始概念之间相差甚远，或者两者之间的关系是间接和不确定的，甚至根本没有相关关系。但有些研究者对这个问题并没有给予足够的重视：随便给某个抽象概念找个指标，然后在后续的研究中就直接使用

这个指标。在社会科学中,类似的错误非常普遍,研究者在从事定量研究中应该注意。让我们举几个这样的例子,比如研究者试图利用邮政、贸易、旅游和交换学生等方面的信息创建一个关于欧洲“社会一体化”指数;或者研究者询问被调查者希望环境改善还是希望赚更多钱,然后根据不同回答将被调查者分为“物质主义者”和“后物质主义者”;再或者研究者测量出不同联邦机构中雇用职员平均雇佣时间,并将雇佣时间长短视作不同机构“制度化”的指标。

读者应当注意我们在以上内容中提出的观点。在社会科学研究中,概念和指标之间有差别是不可避免的。较之特定术语,使用一般术语允许研究者扩大参照系及理论的适用性。比如,我们就可以讨论一般意义上的立法机构而不是那些特定的立法组织,如美国国会或德国联邦议会。或者,当我们希望理论的适用范围更广时,如想讨论一般意义上的“决策机构”而不是“立法机构”(将在下一节详细展开)。科学确实依赖于这样抽象的分类,否则我们就要不断地总结历史。但是,较为抽象及一般性的术语必须在某种程度上与特定的及可测量的概念相关,只有这样才能在此基础上进行实证检验。因此,这两者之间既要相互联系也必须保持适当的距离,这一点研究者需要牢记于心。此外,只有从手头的理论问题出发,才能在此基础上采用更加抽象概念。同时,抽象概念必须有助于我们将那些包含特定指标的研究问题和一般性问题关联起来。此外,采用其他指标的进一步研究也会支持你所选取的指标确实和一般化的概念相关。以上例子中提到的“社会整合”(societal integration)、“后物质主义”(post-materialism)及“制度化”(institutionalization)等抽象概念便可以通过某些特定指标加以测量。在特定指标和抽象概念之间进行切换是必需的,研究者在从事社会科学研究时,必须学会实现两者间的切换。需要提醒研究者注意,在切换时需要十分谨慎:不光理由要合理,也要牢记切换由何处开始。

综上所述,我们并不反对抽象化。在社会科学研究中使用的语言应该尽可能具体和精确。在具体研究中也许只能用一些不能被直接观察的概念,在这种情况下,研究者至少应该选择那些能够观察到结果的概念。举个例子,虽然“智力”这个概念很重要但却难以直接测量,我们

可以用很多智力测试来测量这个概念。又比如，如果必须在“总统制的制度化”及“白宫工作人员的数量”这两个指标中选择的话，选择后者比较明智。虽然“白宫工作人员的数量”和“总统制的制度化”这个一般概念相关，但也不能据此就将一个狭隘的概念和一个广泛的概念等同起来。如果成员数量意味着制度化，我们也可能找到其他制度化的测量指标，这些指标应该和成员数量相似且都可以做为解释变量。以下内容，我们将讨论如何通过增加被解释变量来最大化解释力度。

在描述一个理论时，使用的语言应当尽量具体。试想一下，如果读者花了很长时间去理解理论的含义，该理论的作用就会大打折扣。对理论的描述应当使得读者很容易知道研究者试图表达什么。即使研究者并不亲自从事实证检验工作，也应当花点时间仔细考虑理论可被观察到的结果及研究计划。在这个过程中，虽然较为模糊的表达方式可以降低犯错误的概率，但研究本身的有用性也会降低。总而言之，犯错也比表达不清楚要好。

一般而言，在展现研究的逻辑、论述重要性和描述事件时，我们总是希望看到传神的写作，当然这类作品在社会科学中并不常见。而冗长乏味的文字却不能推动科学的发展。不过一旦研究的目标是因果或描述性推论，且期望从中获得的结论具有一般性的话，研究者使用的语言和思维方式就必须具体和明确。①

① 在采访中掌握最佳提问的规则和设计解释的规则是相同的：要尽可能具体。我们不鼓励直接问那些较为保守的美国白人“你是种族主义者吗”，一般来说，应该这样问：“你是否介意你的女儿嫁给一个黑人？”我们也不能直接问受访者是否了解政治，而是应该问他/她是否知道国务卿和白宫发言人的名字。在访谈中，我们不应该要求受访者去做那些本应该由我们做的工作。比如，最好不要直接问他们对因果效应的估计。但是我们必须问解释变量和被解释变量的测量方法，然后由研究者自己完成对因果效应的估计。我们应该询问事实而不是动机。

当然，这个规则并不是说研究者不需要去询问受访者为什么要做某件事。事实上，对受访者动机的询问是获得研究假设的一个有效途径。受访者自己述说的动机可能就是关于理论可观察的现象。受访者给出的答案应当被理解成其在访谈时所做出的反应，而不一定是正确的答案。如果这类问题有用，那么我们就应当进行研究设计以便获得的回答（不管回答的是正当理由、被修饰过的答案、谎言还是选择性记忆，这些研究者在访谈中随时都能遇到）能作为理论的可观察现象。

3.5.5 规则5:理论的阐述应当具有一定的灵活性

在实现了理论可以被证伪及具体化之后,建构出的理论应当尽可能多地解释现实世界。这条规则和之前讨论的具体化也许互相对立,在某些情况下,甚至发生冲突。实现两者间的平衡需要研究者对它们的差异非常敏感。

例如,如果你的理论适用于所有独立的立法系统,在建构理论时,你就不能让别人觉得你的理论仅适用于德国联邦议会制度。在陈述理论时,研究者并不需要提供该理论的全部证据,但需要对该理论的不确定部分进行合理估计。研究者已经提供了有力证据表明该理论适用于德国联邦议会。我们并不清楚这个理论是否适用于其他国家和地区的立法机构,但也没有证据证明它不适用。如果研究者认为有必要对该理论的一般性进行评价,就需要用该理论进行更多的推论。从这个意义上来说,将进一步的推论表述成一个假设有助于我们理解理论的机构性特征。而正是这种结构特征决定了这个理论是否能帮助我们理解其他立法机构。比如,该理论是否可以用来解释美国参议院制度?和德国联邦议会不同,美国参议院议员的任期并不相同;这个理论是否可以用来理解新罕布什尔州的州议会?众所周知,该州的议员和选民关系较为紧密;或者用来增加对英国国会下院的了解?在那里政党力量比较强大。实现解释力度拓展的重要步骤就是找到理论中适用于不同领域的系统性特征。在这个过程中,研究者也许会犯错,但这也比从一开始就没有严谨地建构理论要好。

我们的这条建议貌似与罗伯特·默顿提出的“中距理论”(theories of the middle-range)存在冲突。如果读者阅读过默顿的著作就会知道并非如此。默顿认为传统社会学中的一些理论,比如帕森(Parson)的“行动理论”就因其阐述得过于宽泛而无法被证伪。在政治科学中,伊斯顿(Easton)的“系统理论”(1965)也属于这种类型(见 Eckstein,1975:90)。“角色设置理论”(theory of role sets)就是一个默顿乐于批评的例子,默顿曾说道:“据我们所知,这个理论和马克思理论、功能分

析理论、社会行为主义、索罗金(Sorokin)或帕森的行为理论等过于宽泛的理论没什么不同。”当然默顿并不是在批判“角色设置理论”(他称之为“中距理论”)本身,他反对的是那些“定位过于宽泛的理论”:几乎所有特定理论或实证观察结果超不出它的解释范围!默顿倾向于“中距理论”,但我们更相信他认同这样的观点:在可被证伪且明确的前提下,理论的建构应该尽可能宽泛,套用我们的术语就是“最大化解释力度”。当然,过于宽泛的理论可能无法被检验,但还是那句话:只要能被检验,理论越宽泛越好。越宽泛,解释力度就越大。

4 确定观测对象

以上章节已经讨论了定量研究和定性研究中科学推论的标准（第1章）、描述性推论（第2章），并明确了因果关系及因果推论的相关概念（第3章）。本书接下来的内容将结合相关文献和具体例子分析定性研究中经常遇到的实际问题。我们在本章中首先讨论研究者如何选择样本及观察值。在从事此项工作时研究者应该非常谨慎，如果稍有偏差，研究就有可能无法获得准确的因果推论，其结论也就失去了意义。在第5章中，我们会讨论导致取样出现偏差甚至无效的原因：只有厘清源头，才能通过调整估计来避免这些错误。本书最后一章（第6章）要解决的问题则是如何根据已有的数据尽可能获得更多观察值。其实，研究者是可以在现有数据中发展出更多观察值的。正如我们在第1章中总结的那样：研究者应当用尽可能多的可观察现象并尽量利用它们去检测理论。

在3.3.2小节中，我们讨论了“有条件独立”概念：一个解释变量的选择与赋值均独立于被解释变量。一旦解释变量的选择标准与被解释变量相关，或者被解释变量是解释变量的原因，那么该假设就不成立了。在基于大样本的定量研究中，为确保该假设成立，研究者往往采用随机方法对解释变量进行选择和赋值，这样可以在一定程度上缓解因果推论中的根本性问题。但随机的方法并不适用于小样本研究。在小样本研究中，研究者通常采用的是有目的地选择观察值（已经在3.3.1小节中讨论）以期获得较为同质的观察单位（unit homogeneity），并由此保证因果推论的有效性。从某种意义上说，有目的地选择观察值是保证因果推论有效性所能采用的“最后一道防线”。

在开始本章的讨论之前，先让我们回顾一下“单位同质性”假设：如果关键解释变量的两个观察值相等的话，那么与之相应的被解释变量的预期值也应当相等。我们通过一个例子来说明该假设的严格版本：打开电源开关就会点亮一只 60 瓦的灯泡，那么如果让另一只同样的灯亮的话，你要做的同样是把开关调制到“开”的状态。在这个例子中，开关的位置是关键解释变量，灯泡的状态（亮或是灭）则是被解释变量。“单位同质性”假设成立的前提是：只要开关处在相同位置，则每盏灯的预期状态都相同。如果将该假设的要求稍微放宽，让其更为合理且易于接受的话，就是“不变效应”假设：尽管变量的等级可能不同，只要解释变量的两个观察值出现相同幅度的变化，那么其导致的因果效应应该相同。拿电源开关的例子来说，假设电源开关有三种设置：一种设置是从“关闭”调到“低亮度”；另一种是从“低亮度”调到“高亮度”。同时我们根据每种状态的功耗来测量被解释变量：“关闭”状态消耗的瓦特数是 0 瓦，“低亮度”是 60 瓦，“高亮度”是 120 瓦。读者应该注意到上述两种调整并不相同，但只要它们导致相同的瓦特数变化，那么“不变效应”假设就成立了：不管是从前者的 0 瓦上升到 60 瓦，还是从后者的 60 瓦上升到 120 瓦。

如果“有条件独立”和“单位同质性”假设都不成立，因果推论就可能存在严重的问题。若在此时，研究者连一个明确的研究设计都没有的话，问题就会愈发严重：无法获得有效的推论结果。一个明确的研究设计是有效因果推论的必要条件。我们在 4.1 节中将讨论在一个不确定的研究设计及上述两个假设都不成立的情况下所造成的样本选择偏差问题。4.2 节则分析随机取样和赋值的局限性。4.3 节讨论的重点放在根据被解释变量有目的地选择样本方法可能带来的不利后果。在该节中，我们将会通过具体例子来说明由此产生的偏差会导致因果推论无效。最后在 4.4 节中，我们将系统地阐述如何通过有目的的取样方法使得“单位同质性”假设成立。本节不仅会告诉读者理想研究设计的基本思路，同时也会阐述在条件不甚具备时，有哪些次优方法可供选择。

本章关于样本及观察值选择的讨论应当在研究中被予以足够的重视。在以下的讨论中会涉及许多术语，所以有必要在开始前就给出它们明确的定义。在很多定性研究中，“案例”(case)一词会经常出现，诸

如“案例研究”(case studies)或者“案例方法”(case method)。但这个词含义并不是十分清晰,它可以指一个单独观察值,也可以被理解成一个研究对象。2.4 节中,“观察值”被定义为解释变量和被解释变量一个单位的测量结果。如果是后者的话,它本身就会包含很多变量,甚至有些时候构成研究的主要内容。

为了便于读者理解,让我们以一个例子来说明上述概念。研究者试图关注“关于印度的案例研究”或者“关于第二次世界大战的案例研究”。如果研究重点在于国家间人口分布或者现代战争的阵亡人数,印度和第二次世界大战都可以被视作单独的观察值。然而如果试图进一步探索其他学者感兴趣的问题,可以从它们中进一步区分出很多观察值,而这些观察值又能够进一步细化为研究单位或变量。比如,研究者试图比较印度各州的选举结果或第二次世界大战中历次战役的结果。上述事例表明在研究设计中,把印度或者第二次世界大战定义为“案例研究”可能会误导读者:两者也许仅仅是定义了一个包含大量观察值的大概界限。

无论研究对象是什么,研究者在选择观察值时都应当选择那些可以获得推论的观察值。因此,我们建议社会科学研究者基于观察值展开研究,而非案例中那些松散的术语概念。不过在定性研究中,更为普遍的做法是研究者首先选定“案例”并将其视作一个观察值,而后当意识到观察值不足时又不得不再将这个案例拆分。

举个例子,如果研究者试图研究在贫穷的民主国家中不同经济增长模式对政治制度的影响。在操作中,他也许会把 1950 年到 1990 年间的印度视作一个案例。这样做的理由是研究者已经预先根据两个变量(经济增长率和政治制度的变革及稳定性)将印度作为研究对象。但随着研究的进一步开展,研究者却发现只能找到极少数的贫穷民主国家:观察值数量不足以做出有效的因果推论。在意识到这个问题之后,他所能做的应当是增加观察值的数量:把印度各邦作为研究单位,同时再把样本时间跨度分成四到五段。通过这种方法,研究者便获得了足以检验理论的样本数量。如果通过这个过程获得的信息足以支持关于印度政治的因果体论,其产生的结论可能完全不同于那些仅基于一个观察值的传统案例研究。

我们认为“观察值”是一个比“案例”更为确切的定义，本章都用“选择观察值”来表达取样过程。不过很多研究者通常是先选择一个包含多个潜在观察值的研究范围，并约定俗成地将这个过程定义为“案例”，所以在讨论这些实际研究时，我们仍然用“选择案例”，而非我们建议的“选择观察值”。

4.1 不明确的研究设计

研究设计是关于研究的计划，指的是研究者通过对模型和数据的讨论并利用手中的证据进行因果推论的方法及过程。定性研究者一般不会先给出明确的研究设计，而是将其隐含在研究过程中的各个阶段里。但是有些研究设计存在不明确的问题，研究者便很难从中获取有价值的因果推论。

不管对于定量研究还是定性研究来说，研究设计不明确的问题比比皆是。但这个问题在这两种研究方法中导致的后果却不尽相同。在定量研究中，该问题导致的后果一般很容易被发现：计算机程序无法估计出结果。①不确定的统计模型使得计算机无法按照预期运行，同时也无法得出任何有意义的结论，类似的问题在定量研究中非常常见。遗憾的是，并没有哪种计算机程序可以自动纠正研究设计的不明确问题。但研究者如果意识到该问题的话，至少有助于时刻警惕随时发现不确定的地方并找到相应的解决办法。在定性研究中，由于研究者掌握大量信息，可以在更大程度上确保研究设计的明确性。

假定搜集信息的目的是检验某个假设的有效性，正确的研究设计应该能提供最大化的解释效力，以便研究者能够区分出所有可能的相关结果。在以下两种情况下，研究设计会出现不明确问题，从而导致无

① 在计量经济学及统计学的相关文献中，“识别”这个概念主要指通过修正模型或搜集其他类型的数据来处理由于不明确的研究设计所导致的问题(Hsiao, 1981; King, 1989:8.1)。

法获得最大化的解释效力：

1. 推论数量大于观测值数量。

2. 如果解释变量中有两个或两个以上的变量在统计上密切相关，那么就存在多重共线性问题(multicollinearity)。这些密切相关的变量可能并不相同，但如果它们可以准确无误地预测彼此的话，就会导致不明确问题。

需要注意的是，上述两种情况及研究设计的非明确性都是针对因果推论来说的。一般来讲，关于汇总历史细节的研究设计不会存在这个问题，除非研究者搜集的观察值都和研究对象无关。如果研究者至少掌握了一定的信息，为了提出一个有趣的研究问题(见2.1.1小节)而进行的数据收集工作一般不会导致研究设计的非明确性。当然，随着研究的推进，这个问题可能会慢慢凸显出来，比如研究者为了检验因果假设而重新概念化既有数据(或者搜集新的数据)。

4.1.1 推论数量大于观察值数量

让我们首先考虑上面提到的第一种情况，即研究中推论的数量大于观察值数量。推论是指根据已知事实去研究未知事实的过程。但从有限的已知信息中获得的推论数量也是有限的。一般来说，从一个信息(或观察值)中无法获得一个以上的独立信息。也就是说，从一个观察值中至多获得一个推论。如果观察值不是彼此独立的话，从 n 个观察值中获得的推论要少于 n 个。在实际研究中，研究者作出一个合理因果推论所要求的观察值数量远不止一个。

在定性案例研究中，读者经常会遇见推论数量超过观察值数量的情况。当然并非所有的定性研究都会有这个问题。在那些将检验理论的观察值被错误地概念化及组织时，研究者才需要特别留意这个问题。我们在下面的内容中将对该问题进行讨论并给出具体解决方法。

假设研究者手头有三个案例，每一个都有对应的两个国家，它们通过彼此合作来建造现代化的尖端武器系统。该研究包含很多有趣且详尽的描述，比如武器系统、国家间的谈判过程及最终合作成果。在研究

过程中,研究者罗列了七个促使各国在首都防卫计划的谈判中达成一致并成功合作的重要决定因素。这些因素可以被用来作为解释变量。同时,研究者也采访了各国的决策者。这些政治领袖也认为这七种因素是重要的解释变量。这样一来,研究者就有七个合理的假设及八个变量的观察值:七个解释变量和一个被解释变量。基于这些信息所作的推论工作却难以避免因果推论的根本性问题。虽然研究者在这个例子中使用了结构性及集中比较等很有价值的方法,但这些方法在这种不明确的研究设计下能够发挥的作用十分有限。

面对这些不明确的解释,我们有时会考虑解释事件发生的其他原因。但这种做法与解释问题应当采取的逻辑恰好相反。这种通过对案例进行全面详细的描述并不会有助于问题的解决。因为在参数数量多于观察值的情况下,对这七个变量的作用考察所基于的样本大小彼此并不会有差别。此时,无论施以多么详尽的描述、采用多么匠心独具的方法,或研究者有多么强的研究能力,都无法获得有效的因果推论。事实上,把所有可能的解释变量都纳入研究范围的努力会把研究推向非明确的深渊。

寻找大量额外的观察值有助于上述问题的解决。当然这意味着更多的时间和资源,这也许会超出研究者的能力范围,或者仅有这三个案例可供研究。我们对此的建议是把重点重新放到某个特定解释变量上而不是关注所有解释变量,去检验该因素在国家行为中产生的作用。例如国际合作项目为什么成功?除此之外,如果不愿意完全改变研究重点,可供研究者选择的另外一个方法是在分析的不同层次上增加观察值数量。比如除了武器系统的合作外,研究者也可以深入考察关于每个系统建立的决策过程。如果决策过程蕴含的额外信息与因果推论显著相关的话,将它纳入考察范围是非常有帮助的。只要对理论的预期和决策效果相吻合,就完全没有必要改变研究目的。经过上述设置,我们就为检验理论增加了额外的观察值或者在数据中拓展出了额外的观察值。后者对于定性研究尤其有用:通过集中于案例的分支部分进一步将一个案例分解为若干观察值。从研究的不同层级增加新的观察值可以使研究者采用多种方式来检验假设。该方法不只对定性研究尤

其有效，同时也可以（在一定程度上）避免非明确性及遗漏变量偏差问题（我们将在本书 5.2 节中讨论这些内容）。对于如何通过研究设计来扩充观察值数量将在本书第 6 章（特别是 6.3 节）集中讨论。

推论数量大于观察值数量的公式化表述

理解这个问题最简单的方式就是通过例子说明。为了便于读者理解，以下内容将尽量采用直观的表达式表述，我们尽量不采用复杂数理推论。但这并不妨碍这些推导过程的一般性。

假设以下表达式包含两个解释变量和一个被解释变量，我们的目的是对这两个解释变量的参数进行推论：

$$E(Y) = X_1\beta_1 + X_2\beta_2 \tag{4.1}$$

但此时我们只有一个观察值可以被用来估计（即 $n = 1$）。这个观察值包括：$X_1 = 3$，$X_2 = 5$，$Y = 35$。同时假设 Y 在这个例子中正好等于它的期望值（这种情况的出现一般纯属偶然，也可能是因为 Y 中没有随机性变化），即 $E(Y) = 35$。但在实际情况中是不可能获得这样的关键信息的（因为 Y 一定带有随机性）。在这种情况下，研究者没有关于期望值的足够信息，也就不可能把参数 β_1 和 β_2 的效果估计出来。

换句话说，此时的目标就是在如下公式中估计出这些参数的系数：

$$\begin{aligned} E(Y) &= X_1\beta_1 + X_2\beta_2 \\ 35 &= 3\beta_1 + 5\beta_2 \end{aligned} \tag{4.2}$$

显而易见，通过上述表达式无法获得唯一解。例如，$\beta_1 = 10$ 和 $\beta_2 = 1$ 的组合可以满足该公式，但 $\beta_1 = 5$ 和 $\beta_2 = 4$ 以及 $\beta_1 = -10$ 和 $\beta_{13} = 13$ 同样可以使得上述公式成立。这种情况的出现会给推论带来困难，因为在有不同参数组合选择的情况下，研究者无法获得关于这两个参数的确切因果效用。在最后一个组合中，参数的方向甚至都发生了变化。此时，研究者无法确定哪一个参数的估计值才是我们需要的，因为上述所有组合均满足方程式。

4.1.2 多重共线性

假设在研究过程中，为了解决观察值不足的问题，研究者采用的方式是由解释变量推测它们的作用，而不是从结果反过来寻找原因。那么通过增加不同研究层级中观察值数量的方法可能会导致另外一个引起不明确性的问题：多重共线性。该术语源自统计学，读者在回归分析的研究中会经常碰到这个词。在本书中，该术语的使用较为宽泛：所有能用一个或多个现有解释变量去准确预测其他解释变量的情况都被称为存在“多重共线性”问题。和统计学研究中的假设不同，在本书中不会涉及共线性假设。

回到前文那个军事合作的例子，该例子存在如下两种假设：(1)较之领土面积相似的国家，领土相差悬殊的国家更有可能成为军事合作伙伴；(2)较之相邻的国家，那些不相邻的国家间成功进行军事合作的概率会更高。这两个假设的解释变量都突出了竞争关系对合作的消极影响。逻辑上看，这两个假设的设置都很合理：现有的基于深度访谈和产业政策的相关文献都可以为其提供支撑。现在假定每个分析单位仅包括两个国家，同时研究者能够收集到的样本非常有限。比如只有两类观察值能搜集到：领土面积不同但相邻的国家和领土面积相当但不相邻的国家。如果观察值（不管是有意设计还是纯属巧合）都属于这两个类别，那么基于这些数据来证实或证伪上述任何一个假设都是不科学的。其原因就在于这两个解释变量之间是高度相关的：如果两个可能成为军事同盟的国家在领土面积上比较相似，两者很有可能是相互接壤的，反之亦然。目前基于数据，研究者无法明确区分出面积和地理位置这两个在概念上完全不同的变量。此时最理想的做法就是继续搜集其他观察值，比如那些领土面积相当且地理位置相邻的同盟国家。如果无法做到，剩下的唯一出路就是在分析的其他层级去搜寻样本了。

即使解决了研究设计的不明确问题，如果观察值数量不足或解释变量之间存在共线性，研究者就依然无法获得确切的因果推论。为了

增加推论的信度，研究者应当尽可能多地增加可以检验理论的现象以最大化已有样本的解释力。当然在实际中，学者去收集数据的时间及资源都很有限，但应当时刻牢记观察值数量要大于推论数量这点。不能一味地围绕一个案例去搜集所有信息，而应当把注意力放在搜集其他类似案例的信息上。对于不明确性的担心也会影响对分析单位的定义：如果研究者找不到其他可用样本，只能凭借某个独特案例展开分析的话，那么就无法获得有效的因果推论。即便研究对象是像共产主义、法国大革命或民主的根源这样的宏大议题，进一步将其分解成大量易处理的分析单位也是一个值得尝试的努力。

另外一个能够有效地实现解释最大化的方法是限制解释变量的数量。运用这种方法，研究者必须时刻警惕遗漏变量偏差问题（见本书 5.2 节中的讨论）。5.3 节则给读者提供了一个好研究的标准：基于的信息很少但解释的问题很多。当然如果可以做到，最好的情形就是只用一个解释变量就能解释大量现象。

在我们看来，一个基于很多信息去解释同样多问题的研究设计意义并不大，但是一个不明确的研究设计却根本无法估计出因果效应。为了避免这个问题，研究者需要为变量搜集更多的观察值。以下内容将用公式表述多重共线性，在这之后我们将详细讨论取样方法及取样偏差问题。

多重共线性的公式化表述

与上节类似，我们同样通过一个例子来帮助读者理解该问题。和上面的公式化表述类似，以下公式虽然简单直观，却具有广泛的实用性。

我们依然采用公式(4.1)中的模型，不过这次研究者搜集了大量的观察值，同时这两个解释变量之间存在完美的线性相关关系。为了理解的便利，我们进一步假设这两个变量的取值相同，即 $X_1 = X_2$。虽然有相同取值，但这两个变量是完全不同的，比如性别和怀孕与否。但是在研究过程中，这两个变量在数值上相同，比如调查时所有的妇女恰好

在孕期。读者可以试想一下，在这种情况下还能估计出这两个变量各自的作用吗？

此时公式(4.1)可以写成：

$$E(Y)=X_1\beta_1+X_2\beta_2 \\ =X_1(\beta_1+\beta_2) \tag{4.3}$$

能够作这样的变换的原因是 X_1 和 X_2 是相等的。显而易见，不管 $E(Y)$ 和 X_1 如何取值，我们都能找到很多 β_1 和 β_2 的组合满足上述公式。比如 $\beta_1=5$ 和 $\beta_2=-20$，$\beta_1=-20$ 和 $\beta_2=5$。在这种情况下，尽管研究者有多于推论数的观察值，但由于多重共线性问题的存在，我们依然无法获得上述参数的唯一估计。

4.2 随机选择的局限

在大样本研究中，研究者往往通过随机取样来避免估计偏差，因为随机法则与研究中的解释变量和被解释变量都不相关。[①]随机取样是一种很强大的方法，它可以自动实现取样过程与所有变量都不相关。换句话说，随机法则使得大样本研究中的取样方法和变量之间相关的可能性变得微乎其微。因此由选择性而造成的偏差问题就不存在了。在一个充满不确定性与干扰因素的世界中，随机取样无疑是社会科学家的福音。如果出于某些限制不得不放弃该方法的话，研究者必须非常谨慎地实施，而这在政治科学领域中非常常见。

① 在这里，读者不能将随机性和任意性相混淆。这里，随机选择的意思是每一个单位都有同样的概率被选进样本中，同时前后选择都是彼此独立的。我们经常看到从帽子里抽出写有人名的字条然后再放入一个新字条，这就是一个很简单的随机选择的例子。需要指出的是，所有的随机选择都需要特定的概率过程。

对于社会科学来说,控制实验并不是被研究者经常采用的研究手段。[①]但控制实验中的思想却有助于我们理解那些非实验性的研究设计。对于实验方法来说,最有效的方式通常是把随机取样和基于大样本对解释变量(或实验)的随机赋值结合在一起。即使实验解决不了因果推论的根本问题,但基于该方法的研究者可以自主取样(而非从真实世界中搜集)并给每个观察值是否被介入赋值。因此,实验有两大优势:能够控制对观察值的选择及能对解释变量赋值。在实际操作中,由于客观条件的限制,研究者一般不是随机取样而是借助便利的方式搜集观察值,例如招募大学二年级的学生。不过为了讨论的便利,我们只关注最理想的情况。本节首先讨论对观察值的选择,而对解释变量的赋值我们将在第 5 章结尾讨论。

在定性研究和定量研究中,由于案例无法被清晰地选择出来,因此随机取样并不是在任何情况下都适用。比如,针对美国外交精英的随机取样就很难,因为没有一个像国会选区目录那样明确的精英名单。研究者也许会从很多渠道获取名单,这就会导致取样偏差。比如,研究者将取样范围限于那些接受过政府外交政策咨询的公民,这些人当然可以被视作外交精英。但如果研究者感兴趣的是社会背景对政策偏好的作用的话,那么搜集到的人往往都处于社会的上层同时偏向于支持政府决策,这就会导致估计结果出现偏差。此外,研究经费的限制往往也使得我们无法进行整体的随机取样,在实际情况中往往是将研究对象局限在某个地理范围,这也可能导致偏差。

即使随机变量可行,也并不意味着它总是最好的选择。定性研究者有时会故意避免使用随机选择方法,因为这样做可能会使一些重要的案例被错过。(如果不研究法国大革命,那对人类历史中革命的理解不就失去意义了吗?)事实上,如果观察值数量有限的话,随机取样并非明智之选。它并不能解决样本选择偏差问题,甚至还不如其他取样方法。正因为这点,对于定量研究者鼓吹的随机性更科学的论点,很多定

① 相关例子可以参考 Roth(1988)、Lyengar & Kinder(1987)、Fiorina & Plott(1978)、Plott & Levine(1978)及 Palfrey(1991)。

性研究者并不以为然并认为有误导性。其实只要一个非常简单的模型就能证明在小样本研究中,随机选择会导致严重的偏差。

假设被解释变量有三个取值:高、中、低。在这三个值之间我们选取两个进行分析,也就是说观察值数量为 2。首先确定选择法则:1 表示该观察值被选取,否则是 0。那么只可能有三个选择法则:(1, 1, 0)(表示取值为“高”和“中”的观察值被选取,“低”没有被选取)、(0, 1, 1)、(1, 0, 1)。这三个选择法则中,只有第三种才与被解释变量不相关。① 随机选择观察值就相当于在这三个法则中随机选取一个,我们会发现在这样的小样本研究中,采用随机选择的话,其造成的偏差概率竟高达 2/3! 这个事例就告诉读者,在样本较小的情况下应当基于已经积累的知识去选取被解释变量,只有这样才能提高选到第三种组合的概率从而避免偏差。

相比较定量研究,定性研究很少采用随机取样方法。这也使得该领域的研究者必须非常谨慎以确保其取样标准不会造成类似的后果。举个例子,研究者对第二次世界大战后苏联统治下的东欧天主教国家感兴趣,这些国家包括捷克斯洛伐克、匈牙利和波兰。该研究者发现,20 世纪 70 年代至 80 年代期间上述国家在政治上存在显著差异:波兰发生了组织完善的反政府运动(波兰团结工会),捷克斯洛伐克国内也一度活跃着较小规模的知识分子群体(“七七宪章”组织),但相比之下,匈牙利却从未出现过类似的全国运动,研究者试图解释的正是这种差异。

为了探索反政府运动,研究者必须对报纸信息、近期共产党解密文件及运动参与者的访谈内容进行分析,因此语言的驾驭能力就显得非常重要了。此外,针对当代东欧国家研究的另外一个困难是,每个国家都得花上研究者一年的时间。因此最可行的方案是只选择两个国家进行研究。幸运的是,由于某些与研究计划无关的原因,研究者掌握捷克语和波兰语。她就决定选择捷克斯洛伐克的“七七宪章”组织和波兰的

① (1, 1, 0)的选择规则忽略了最低值(即“低”这个单位),(0, 1, 1)则忽略了最高值(即“高”这个单位),只有忽略了“中”这个单位的第三种规则(1, 0, 1)才与被解释变量是不相关的。

团结工会作为研究对象。显而易见，这并不是一个随机取样过程，但只要选择这两个国家的原因与被解释变量是无关的就可以。然而她所依据的选择标准（语言学知识）却的的确确和被解释变量相关，这就会导致样本选择偏差。在这种情况下，非随机选择方法也许更为合适。当然，其前提是选择标准不是出于语言上的考虑。

为了避免这种偏差，研究者没有利用自身的捷克语优势，而开始学习匈牙利语。但这依然不是个好方法！在我们看来，更现实的策略是利用对样本选择偏差的认识来判断偏差方向，尽量去修正偏差从而证明从中获得的结论是合理的。首先，她需要意识到这种系统化的选择方法降低了解释变量的变化幅度，这会导致至少从平均值上低估了真实的因果效应（当然研究中的其他问题可能也会改变这个结果）。

其次，即使没有选择匈牙利，她也应该对这个国家进行足够的研究，以弄清楚样本选择偏差是否会影响针对现有解释变量的假设。比如，根据捷克斯洛伐克和波兰的情况，作者认为那些大规模的反政府运动都是在政府持有较为宽容的态度时酝酿的，如果政府采取强力镇压的态度则形成的概率较低。但针对匈牙利的研究就会发现，虽然该国政府在所有东欧共产主义政权中是最为宽容的，那里却没有发生过任何大规模的反政府运动。这就告诉我们应该尽可能地扩大考察范围，以避免样本选择偏差。即使不能对所有观察值进行研究，但掌握更多额外的信息至少能够在一定程度上降低这种偏差。一个很有用的策略就是在已有两个案例的详尽考察之外，再依靠二手数据不那么详尽地考察一些其他案例作为补充。如果详尽的案例分析可以获得明确的因果推论的话，那么利用更多的样本去搜集那些重要变量的信息就变得相对容易些（更详尽的讨论见 4.3 节）。另一种可行的方法是将已有两个案例中的大量信息重新整理，以期获得更多用来检验理论的样本。比如，假定政府通过镇压手段可以有效抑制反政府运动的理论是正确的，那么在一个国家之内，我们就会预期运动很难在秘密警察势力强大且有效率的地区兴起，反之亦然。

4.3 样本选择偏差

在实际研究中，我们应该如何选择观察值呢？如果你打算采访市政官员，哪些人才是理想的采访对象？如果想对主要战争进行比较研究，应该选取哪些战役？如果对总统否决权感兴趣，是从所有此类案例还是第二次世界大战以来的案例中随机选择，还是只选那些被国会否决的案例？在研究设计的初始阶段会遇到的问题就是：应当选择哪些案例（或者更确切地说，哪些观察值）来进行研究？对于定性研究来说该问题尤其关键，往往会影响研究的深入程度及结果的可靠性。

在4.2节中我们已经说过，随机选择一般不适合小样本研究，但放弃随机选择也意味着偏差发生的概率会显著提高。研究者易犯的最典型的错误是，在获得可靠的研究结果前，内心就已经有了预期结果（能证实最心仪的假设），然后再结合被解释变量和解释变量去收集能够支持理想结论的观察值。举个例子，如果认为美国在第三世界的投资是这些国家内部动乱的主要原因，研究者也许会有意识地选择那些接受美方大量投资并曾发生大规模动乱的国家和那些既未接受美国投资也没有动乱的国家作为研究对象。我们当然知道观察值还有其他组合形式（比如大投资无动乱或小投资大动乱），只是这些观察值往往被研究者“巧妙地”忽略掉了。当然，在实际研究中的样本选择偏差并不都如此明显。对于定性研究来说，样本选择的标准并不是非常清晰，此时如果研究者不会刻意评估潜在偏差的话，就会导致偏差乘虚而入。①

① 这个例子能够很好地说明为什么科学是独特的。如果研究者为了支持其结论而引入偏差，这种做法就不像一个社会科学家所为。但很多学者表现得如同那些在辩论中捍卫自己立场的政治家一样：总是选择那些能够证明自己观点的证据。在研究过程中，研究者应当尽量去搜集所有的观察值，如果一定要从中选择，应当选择那些重要的观察值，而不是那些能够支持我们观点的观察值。

4.3.1 根据被解释变量选择样本

大样本研究的随机选择方法允许我们忽略取样标准与其他变量间的关系。一旦放弃随机选择,研究者就必须考虑这种相关性。在此,我们提出一个公认的基本结论:样本选择过程应当使得被解释变量有发生变化的可能。有些读者可能觉得这根本不值一提:如果被解释变量没有变化的话,那如何去解释它的变化呢?可是很多研究的确犯了"被解释变量没有变化"的错误。比如一些试图解释战争或者革命爆发的研究,其所依赖的样本确是战争和革命本身。还有一些研究试图通过采访没有投票的选民去解释选举结果。①

在本书第 1 章中我们曾经提到,好的社会科学家经常研究一些反常现象。所以研究者,尤其是定性研究者,会刻意选择那些众所周知但结果却比较奇怪的案例,例如,有的研究者研究 18 世纪法国大革命及发生在 20 世纪中国和法国的革命(Skocpol, 1979)。选取这三个样本说明作者是根据被解释变量选择观察值的,这可能导致潜在的样本选择偏误问题。如果研究者仅根据被解释的特定取值来选取样本而忽略了其他取值的话,就无法获得其背后的真正原因。针对这个问题,西达·斯考切波(Theda Skocpol, 1979)在其研究中进一步增加了 17 世纪的英国、19 世纪的普鲁士/德国及 19 世纪的日本作为所谓的"革命中那些关键时刻"来部分地解决该问题。她将这些观察值当作"控制案例"(control cases)。同时相对于主要案例,作者也并没有将太多精力放在"控制案例"上。根据被解释变量选择观察值所造成的偏差并不意味着我们在研究设计中不考虑它的取值,而是提醒我们注意这种方法可能会带来的潜在风险并去寻找解决之道。我们将在本章其后的部分及第 6 章中就该点展开进一步讨论。

根据被解释变量选择观察值产生偏差的程度也有较为温和但更为

① 一些研究,特别作为某个大型研究一部分的课题,被解释变量被有意设置成不变。这可能是为了有利于因果推论的作出。但在本节中,我们暂不考虑这种可能性。对此的解释见本书 4.4 节。

普遍的版本。例如,在一些研究设计中,被解释变量被允许有偏差,但偏差的幅度受到一定限制。也就是说,在这个版本中,研究者可以控制被解释因变量的取值变化范围。在这种情况下,研究者能够发现导致被解释变量变化的原因,但其推论依然会有偏差的可能。一般来讲,任何与被解释变量相关的取样标准都会降低因果效应的估计值大小(参见 Achen, 1986; King, 1989:第 9 章)。在定量研究中,这种偏差意味着因果推论的估计值低于实际值并向零趋近。而在定性研究中,该问题则意味着实际因果效应比研究者预期的要大(除非研究者意识到偏差的存在并作出相应地调整)。在实际研究中,如果已经知道样本选择偏差的存在,但实在找不到更合适的观察值来避免这个问题的话,至少研究者可以从推论结果中获得真实效应的下限,但究竟在多大程度上低估了真实效应则取决于偏差的程度(即取样标准与被解释变量的相关程度)。因此,如果实在找不到关于偏差存在的证据,通过以上的讨论,读者应当至少对该问题有所了解。

以上谈的是一般的例子,对于该问题的极端例子——被解释变量没有任何变化——并不难处理。既然在这种情况下无法获得因果效应,就尽量避免它!不过如果取样标准和被解释变量相关的话,情况就不同了。这种情况导致的样本选择偏差就不容易解决,因为研究者不一定能搜集到所有需要的观察值。幸运的是,这种偏差并不是毁灭性的。虽然偏差不可避免,但研究者还是可以在一定程度上对其作出预期并作出处理。我们用以下几个例子来说明这一点。

根据被解释变量来选取观察值因为这可能发生样本选择偏差。虽然这难以避免,但针对此的讨论却有助于读者理解该问题。图 4.1 以图表的方式呈现了这个问题,说明即使在偏差情况下,依然可以获得有用信息。图的散点表示一个观察值(比如一个人)。其中横轴代表解释变量——在商学院进修会计课程的数目,纵轴则代表被解释变量——第一份全职工作的起薪(单位是万美元)。我们将散点拟合获得图中的回归实线。该拟合线就表示两个变量间的关系。从中我们可以看出,平均来说,每多修一门课程意味着起薪增加 10 000 美元。实线周围分布的散点表示回归线不能完美拟合的那些学生的情况。这些点和实线

间的垂直距离是预测偏差(给定被解释变量的取值),也是关于该偏差的最小估计。

现在让我们假设一个即将被学院录取的学生想了解他/她毕业后的起薪。由于没有取样的专门训练,这个学生只选择了那些在首份工作中表现出色的样本。也就是说,他/她的研究对象是那些毕业后就找到自己心仪工作的应届毕业生。这样做也许是因为他/她认为提高薪酬的方法就是研究那些高收入的学生,这当然是靠不住的。简单地说,假设他/她只选取起薪超过 100 000 美元的毕业生作为样本。在图 4.1 中,我们用一条 $Y=10$ 的水平实线表示该选择方法。也就是说,只有那些收入在该线之上的样本才被包含在研究中。如果基于这些散点估计出拟合线(虚线),我们就会发现,由于样本选取偏差导致了拟合线比真实拟合线的斜率小:这个学生便错误地认为多修一节课程只能带来 5 000 美元的额外收入。

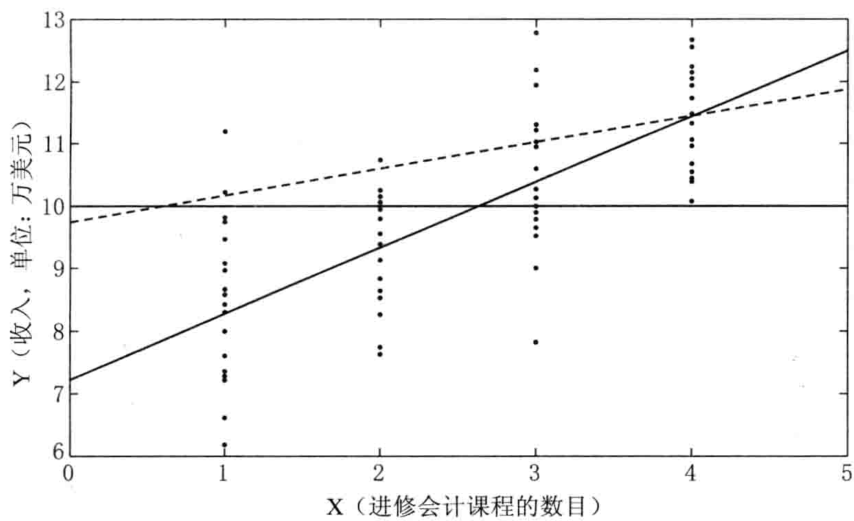


图 4.1 样本选择偏差

上面的例子就告诉我们根据被解释变量选择观察值时,因果效应可能会被低估。幸运的是,该学生完全有能力解决这个问题。假设他/她从商学院毕业后在业界工作几年厌倦了赚钱,于是攻读社会科学研究生并学习了取样知识。由于要忙于应付考试,没法完全重做之前的

估计工作,但此时他/她至少知道修课带来的起薪增长要高于原先估计的 5000 美元。现在他/她意识到进修会计课程有助于提高收入,但此时他/她已经是大学里的研究生了并且决定不会去实现收入最大化了(谁会带着这样的想法去读研究生呢)。不过他/她也许会庆幸自己在价值观改变之前并不知道样本选择偏差问题。

4.3.1.1 由研究者导致的样本选择偏差

上面讨论的问题在定性研究中非常常见(Geddes, 1990)。如果研究者按照被解释变量选取观察值的话,像案例选取这种看上去不会有大问题的研究过程可能正是产生问题的原因。举个例子,假设研究者试图研究近年来总统参与重大外交决策程度的决定因素。当然由于一些决策过程是保密的,研究者只能基于那些已经公开了的信息进行研究。这样的研究设计方案就可能有问题:取样标准(资料的获取)也许只和那些总统参与程度相对较低的决策过程相关(被解释变量)。因为如果决策会议的保密程度越高,总统参与程度可能会越深。但该类会议中的决策过程恰恰是研究者无法获取的。因此,根据信息可得来选取样本可能导致总统参与程度低的那些会议被放大,进而影响对总统参与程度决定因素作用的推论。

上面那个商学院的例子也可以帮助读者了解定性研究中样本选择偏差及其带来的后果。在定性研究中,这类问题是不可避免的。让我们回到总统参与的研究,如果还想知道在那些可能导致军事威胁的事件中,总统参与度是否更高?在分析了大约 24 个观察值之后,该研究者发现现有证据表明,总统参与程度和军事威胁间的关系确实是存在的,不过作用非常不明显。我们有理由相信这可能由样本选择所致。为了评估这种偏差,首先需要尽可能多地收集总统曾经实施或公之于众的外交政策,无论是否完全知道决策过程。然后将搜集的信息绘制成一张表格,使用这种表格能让我们避免一个可能导致样本选择偏差的来源:决策过程保密的程度有可能和武装威胁相关。接下来,我们需要再做一个新的表格,该表格将不包含秘密行动和那些没有付诸实施的行动。但在第二张表中包括的事件则有详细的决策过程。接下来要

做的就是比较这两张表的内容，弄清楚（正如我们所怀疑的那样）会议公开与否是否和实施武力或武力威胁的决策内容相关。如果发现这种相关性存在，那么研究者就有理由认为这两者间的关系可能比原先分析的更强。

根据研究议题收集资料要求研究者必须走访某些特定区域。在比较政治研究中，这同样会导致样本选择偏差。由于一些国家的政府限制外国学者在该国进行学术调研，在这种情况下，研究者的取样就会受到限制。更糟糕的情况是那些限制研究的国家和地区正巧又与研究的被解释变量相关。举个例子，如果研究者试图通过分析反对派的策略来解释威权政体的自由化，那么这种研究得到的结论就很可能存在偏差。尤其是当研究者只能进入当地政府允许的区域，基于这些区域的研究出现偏差的可能性会更大。因为这个区域提供的信息有可能与被解释变量"自由化"相关。当然，我们并不提倡学者到那些不友好的地区进行秘密研究，但如果实在无法获得当地数据，研究者就应当想方设法去寻找其他信息源。及早认识到该问题可以使我们及时修改原先的研究设计，并将视野拓展到其他地区以规避这种进入限制问题。如果这种方法依然不可行，研究也可以对偏差展开分析以了解结论可能存在哪些问题，这样做至少可以部分地改善偏差带来的问题。换句话说，如果样本选择偏差无法避免，研究者此时就应当着手弄清偏差方向，如果有可能就进一步评估其严重程度，然后基于这些认识修正之前的估计结果。

样本选择偏差是研究者经常会遇到的问题。让我们再看几个例子。迈克尔·波特（Michael Porter）在他的书中用"竞争优势"（competitive advantage）来研究当代行业及企业的成功秘诀。为了进行这项研究，他设计了一个大型计划：在十个国家中挑选了十个企业作为研究对象。其取样依据按照他的表述是"那些已经在本行业拥竞争优势的企业，及韩国和新加坡两国中已经有迹象表明即将获得优势的企业"（Porter，1990：22）。然而在研究过程中，波特也采用了根据被解释变量选取观察值的方法，这使得他考察的样本的被解释变量取值近乎相同。这样做的后果是，无论是他本人还是试图这样做的其他研究者，通

过只用这十个国家的数据去解释企业成功的原因，获得的因果推论都将面临严重的偏差。

但是为了找到竞争优势背后的背景及政策，波特还采用了穆勒创造的“契合法”(method of agreement)。该方法当然是个不错的尝试，通过总结不同国家间的相同点，他发展出一套关于竞争优势来源的假设。但单就研究设计来说，借助这个方法来获得可靠的因果效应是不可能的。

除此之外，波特研究中更严重的失误是其逻辑缺陷：缺少对照组(即解释变量的其他取值)。这导致我们无法确定竞争优势是否由这些因素导致。这样的话，他就无法确定那些他认为会带来竞争优势的因素是否在其他案例中导致竞争失败。波特的这部著作在学术界引起的争论很大。因为他的论点都是基于那些获得成功的案例之上，但没有同时考虑成功和失败(也就是说他是根据被解释变量选取样本)。因此，这些证据无法证明或证伪他提出的假设。①

除了波特的这个例子，在外交政策研究中对于“威慑”的讨论也可以用来考察样本选择偏差。“威慑”的定义是“威胁对方，引诱其按照自己的计划行动”(Achen & Snidal, 1989：151)。该领域的学者经常研究所谓的“严重危机”，即在政治斡旋、释放信号及采取行动的初期阶段，此时威慑还没有发生效应。由于这样的阶段可以给研究者提供丰富的信息，正如波特强调的竞争优势一样，研究者可能会认为这是最精彩的部分并花大力气描述。这就不可避免地使得研究者有从结果去推测原因的倾向，从而使得推论产生偏差(假设后来研究者也没有做任何修正的话)。按照这样的研究设计，那些在危机初期威慑政策就奏效的案例就被系统性地排除出考察范围。上面的错误，正像阿肯(Achen)和斯尼德尔(Snidal)所说，“如果错误地将这些案例用于对威慑政策的

① 波特声称在他的书中提到过很多失败国家的例子，但事实上这些案例都是经过筛选后才被纳入他的分析中的。同时，他对这些国家的分析方法也不同于针对那十个成功国家的分析方法，更多的是轶事般的叙述。如果只是非系统性地挑选那些有力的证据，研究者就很容易自欺欺人地认为两个因素间存在因果关系。在这里我们并不是去评论波特的假设是否正确，只是指出如果要获得书中的结论，研究者必须更加系统地收集信息。

估计，那么就会造成样本选择偏差，这和政策评估研究中易犯的错误类似”(Achen & Snidal, 1989:162)。

4.3.1.2 由于受到真实世界的影响而导致的样本选择偏差

如果用总体代替样本进行因果推论，是否能够避免偏差发生呢？有些读者也许会这样想，其实并非一定如此。举个例子，假设我们想评估自由党在纽约州受选民支持的程度，其被解释变量是自由党在该州候选人获得选票的比例。在真实世界中，自由党根本就不会在那些他们认为没有希望赢得选举的地区(实际情况也是这样)选取指定候选人。此时，即使研究者考虑了所有自由党指定了候选人的选区，依然会存在估计偏差。在这个例子中，样本选择过程其实成为选举过程的一部分了，而后者正是研究对象，其导致的后果和研究者出于自身原因导致的偏差一样。

取样过程与被解释变量相关时偏差就会出现，如何避免该问题对于那些以历史记录作为证据来源的学者来说尤为困难。当然，该问题对任何社会科学领域来说都具有挑战性。之所以非常困难是因为研究者并不清楚其依据的历史证据为什么被记录下来。因此，了解这些信息的产生过程就非常重要了。让我们举一个其他领域的例子以便读者理解，有些文明擅长石雕艺术，而另一些文明则擅长木雕艺术品。如果时代久远的话，石雕可以保留下来，但是木雕都腐烂了。这就使得欧洲一些艺术研究者低估了早期非洲雕塑作品的质量及精致程度。之所以低估，正是因为非洲艺术品都是以木头为载体的。而“历史”却选择性地保存了石头雕塑但淘汰了木头雕塑。与此类似，细心的学者就必须常常审视手中的证据，评估是否存在样本选择偏差：哪些信息事件会被留存，而哪些信息则有可能被淘汰。

让我们再举个例子，社会科学家经常以研究终点作为其出发点。比方说针对现代国家中的组织变迁。研究者发现早期欧洲(大约1500年)存在过各种各样的政权组织形式，但到了后来(1900年左右)几乎清一色的都是民族国家。研究者应该做的是从1500年的政权组织形式入手，根据有限的变量解释后来的政权形式。根据这种研究顺序，我

们就会发现很多国家在历史中消失了:有些因为战败,有些因为被其他国家吞并,剩下的则幸存下来。根据这种方法进行归类并获得需要考察的被解释变量:哪些国家在1900年依然存在并成为民族国家?如果不存在,也可以知道它们消亡的时间。

即使如此,还是有很多历史研究者会不经意地犯类似错误。根据查尔斯·蒂利(Charles Tilly, 1975:15)的观察,很多学者采用一种叫"回溯性"(retrospective)的方式作为研究起点——选取"小部分在19世纪和20世纪依然存在的西欧国家进行比较"。不幸的是,这些研究者继而发现,"在激烈的国家竞争中,只有英格兰、法国、西班牙幸存下来,剩下的国家都惨遭淘汰"。1500年的欧洲有大约500个独立的国家,到了1900年只有25个。德国在1500年的时候还不存在,甚至到1800年该国仍未形成。为了揭示出国家的形成过程,研究者比较了法国、德国、西班牙、比利时及英国的历史(也许还有其他欧洲现代国家),借此获得启发。这样一来,整个研究重点就偏向了那些幸存下来的国家了,而它们并不是样本全部,甚至非常特殊。

上述取样过程仅仅基于被解释变量的一个取值:那些在1900年仍然存在的国家。这种取样方法会导致研究出现偏差。一般情况下,该偏差会弱化解释变量的作用大小,这些解释变量本来可以解释为什么有些国家幸存而另一些国家则惨遭淘汰。蒂利与他的同事(1975)意识到该问题的存在,于是在研究设计时放弃使用回溯方式,而采用前瞻(prospective)方式。不过如果这种调整不可行,或者还想为进一步研究准备证据的话,他们也可以对现有基于回溯方式搜集到的信息进行重新分析。从中可以得出现有推论其实低估了真实作用。他们应当意识到即使取样标准没有问题,推论依然有高估或低估真实作用的可能。根据错误的回溯研究,其中样本的选取依据是被解释变量,从平均上来说因果效应一般是被低估的。

4.3.2 根据解释变量选择样本

如果依据关键解释变量选择样本,就可以避免因果推论中的偏差。

其原因在于，研究者没有取样时就确定好研究结果，也就是说，研究者对被解释变量的取值没有施加限制。通过限定关键解释变量的取值范围，可以提高结论的确定性并在防止其太过笼统的同时避免偏差。根据某个变量取值去取样也就实现了对该变量的控制。因此，即使该变量和被解释变量相关，也不必担心推论偏差的问题。①换句话说，在与被解释变量相关的解释变量取样时，只要在分析过程中控制了该变量，就不必担心偏差问题。

读者可以再看一下图 4.1，从中我们不难看出，根据解释变量选择观察值并不会导致偏差产生。例如，如果将所有解释变量等于 1 的观察值都剔除掉，那么该图的整个逻辑都不会发生变化，同时根据剩下的散点得到的拟合实线也不会变。当然，如果和推论有关的观察值和信息量减少，该拟合线的确定性会有所下降，不过一般情况下并不会产生偏差。②

通过以上讨论读者可以知道，根据关键解释变量取样是不会带来偏差的。同样，根据控制变量取样也不会产生偏差（一般来说，控制变量在因果关系上都是先于关键解释变量的，这是所有控制变量的特点）。实验法基本上都是根据解释变量进行样本选择的，研究者将其分成几个种类单位（例如施加药物与否），而后再观察对应的被解释变量变化（病人健康状况是否改善）。实验方法无法根据被解释变量选择样本，因为被解释变量（病人的健康状况）是否变化在实验结束结束之前是不知道的。但确实存在按照被解释变量取样的可能：根据对病人健康状况改变的预期来决定实验方案。这种不谨慎的研究设计会导致从中获得的推论存在问题。

再举一个例子，我们想研究种族歧视对黑人儿童在校成绩的影响。一般来说，研究者应该选择几所种族歧视不严重的学校和几所种族歧

① 总体而言，在控制住解释变量之后研究者再根据被解释变量选择样本，就会导致样本选择偏差。这是因为选择方法本身也是解释变量之一，所以它需要被控制住以避免偏差的发生。

② 如果解释变量的变化范围在取样过程中受到限制，所获推论的确定性同样会降低（见本书 6.2 节）。

视非常严重的学校作为样本。尽管选择标准可能和被解释变量相关(黑人儿童在歧视越严重的学校,成绩往往越差),一旦将解释变量的影响考虑进来,这个选择标准就不再与被解释变量相关了,因为选择样本的标准是由解释变量的一个取值所决定的。

研究者也可以根据与研究无关(也与被解释变量无关)的解释变量来选择样本,这样推论偏差也不会产生。比如,为了研究种族歧视对成绩的影响,研究者选择所有以字母"A"打头的学校作为考察对象。虽然我们并不推荐这种取样方法,但这样做确实可以避免推论结果产生偏差。这个方法成立的前提是该无关变量不是与被解释变量相关的某一变量的代理变量。

在根据无关变量进行取样的同时如果能够结合二手数据分析,可对研究起到极大的帮助。让我们看一个例子:研究者试图探索政变成功的原因并作出如下假设:与文职人员相比,军人领导的政变获得成功的概率要更高。在检索文献的时候,研究者发现了一个关于政变计划的研究,该研究选取案例的标准是该国在政变发生前的官僚化程度。虽然该程度和本研究关注的问题无关,但这个二手资料是可以被利用的。为了谨慎起见,我们建议在比较文职人员和军人对政变的影响时,官僚化程度应当被视作控制变量。在实际操作中,这个不难做到:研究者可以按照官僚程度的高低区分出两种样本,然后分别在其中考察两种人员组织政变获得成功概率的差别。这种方式可以避免样本选择偏差问题,同时其因果效应也能揭示一些与取样过程相关的信息。

4.3.3 其他类型的样本选择偏差

以上例子根据与被解释变量相关的规则或者控制住解释变量之后与被解释变量相关的规则选择样本的话,往往会产生偏差。该类型的偏差在定量研究和定性研究中都是非常常见的问题,并会导致对因果效应的估计值小于真实值。除此之外,还有一种样本选择偏差也值得读者关注。与之前相反,该类型偏差在某些情况下会导致高估因果效应。

假设某些变量的因果效应是随着样本变化的，这种情况到目前为止并没有受到足够的关注，但仍然是可能发生的。本书 3.1 节定义了一个研究单位的因果效应并允许其大小在样本中有差异。举个例子，假设我们试图研究贫困对拉丁美洲国家政治暴力的作用。对于该地区的一些国家，这两者间确实存在很强的相关关系，尤其是那些近期刚发生过政治暴力的国家。在这个例子中，因果效应大小在不同国家(研究单位)中是不同的。如果取样法则与效应的大小相关的话，对于平均效应的估计就会产生偏差。研究者如果仅仅关注那些近期发生过政治暴力的国家，却把基于这些样本的研究结论强加到所有拉美国家，就会导致这个研究中的因果效应被高估。也就是说，如果仅选择那些因果效应比较强的样本作为分析单位，然后依此获取整体作用的话，我们获得的结果就是真实值的上限。同样，如果只基于那些效应较弱的样本，估计值就是真实值的下限。

4.4 有目的地选择观察值

在政治学科学领域的研究中，研究者通常不控制解释变量的取值：很多时候解释变量是由“自然”和“历史”赋值，并不为研究者所掌控。因此，在实际研究中，我们能做的只是如何选取案例和观察值。正如 4.2 节所讨论的，如果观察值有限，我们一般不建议读者采取随机取样方法，有目的地取样此时才是有效的方法。

有目的地选择观察值意味着研究者已经对某些相关变量的取值有所了解。此时，随机取样方法已不再被考虑。对于这种方法，我们的建议是根据解释变量去选取样本，然后获得其对应的被解释变量。在实际操作中，研究者经常遇到很多变量取值缺失的情况。这种情况在取样之前就可能发生，它会在不经意间造成样本选择偏差：人为地使结论偏向预先的假设。在接下来的内容中，我们将向读者介绍几种有目的地选择观察值的方法。

4.4.1 根据解释变量选择观察值

如前所述，最佳“有目的”的研究设计是使得解释变量（以及所有控制变量）的变化独立于被解释变量。相应的被解释变量的取值只有在实际研究过程中才能知道，并通过检验解释变量的某一取值所导致的被解释变量取值分布的不同来获得最初因果推论。

让我们看下面这个例子，假设要研究冷战期间军备控制条约对美苏两国军备采购的影响。在这个研究问题中，针对一个国家特定武器系统的军控条约就是该研究的关键因果变量。研究者可以选择一组武器类型：有些受到条约限制，有些则不受限制，它们都根据解释变量的变化而变化。被解释变量可以是军备采购的变动率，这次我们不根据它来选择样本。此外，如果这两组观察值在其他控制变量上能匹配且有效解决了内生性问题的话，上述研究设计就可以对军控条约的因果效应作出有效推论。

很多因素都可以影响本解释变量，但研究者一般只对一个感兴趣。对于这些变量，研究者需要在分析中加以控制。杰克·斯奈德（Jack Snyder，1991）的研究给我们提供了一个很好的参考。斯奈德挑选了一些他称为“主要权力角逐者”的国家，以便研究其“过度扩张”的程度（被解释变量）。除了他感兴趣的因素外，很多文献都认为军事力量也是扩张的重要解释因素。斯奈德的研究兴趣并非军事力量，在研究中他将该变量视作控制变量：仅选择那些军事强国作为研究对象。这样做的结果就使得军事力量的作用被控制住了，作者便可以集中于自己感兴趣的解释变量，同时也不必担心遗漏变量造成的估计偏差。除此之外，斯奈德研究出众的一点是其工作的探索性质：作者在研究开始之前并没有完全确定所有的解释变量（Snyder，1991：61—65）。这种没有确定目标的研究设计能为研究者带来意想不到的发现，当然也意味着临时碰到的问题可能没有有效的答案，同时解释变量的变化范围也许不如预想的那么大。需要指出的是，除了关键解释变量之外，斯奈德在检验理论的时候并没有使用其他数据。

我们一直强调研究者应当“根据解释变量，而非被解释变量去选择观察值”。读者需要知道的是，这个建议在定性研究中有时不太现实。在收集数据的过程中，研究者有时不得不考虑被解释变量的值，或者现有数据已经包括了这些值。这些并不是致命的问题，但如果研究者还想由此获得因果效应的话，就必须格外谨慎，否则很可能出现偏差。

4.4.2 选择被解释变量一定范围内的观察值

还有一种取样方法是选取被解释变量一定范围内的观察值。有时我们也许对一些行为感兴趣并希望解释其背后的原因。在这种“回溯性研究”设计中(类似在流行病学中的“病例对照”研究)，研究者通常会选取那些被解释变量中取值极高或者极低的观察值。正如我们一再强调的，研究者也许会从此类研究设计中获得因果推论，却无法作出关于被解释变量的描述性推论。此外，由于系统性数据的缺乏及潜在非线性问题的存在，这种取样方法可能无法获得有效的因果推论。

一般来说，被解释变量的极端值也往往对应着解释变量的极端值。因此，这种回溯性研究可能有助于我们获取作出因果推论的信息。但如果研究者希望获得一个有意义的因果推论的话，在选择样本的时候就不能考虑解释变量的取值。也就是说，研究者绝对不能只搜集那些支持自己假设的数据，而应该去搜集那些能够代表整体的数据。如果已经发现解释变量和被解释变量在取极端值时存在对应关系，此时研究者就应当根据解释变量来取样并基于此检验假设。在研究开始之前，研究结果应当具有不确定性，否则基于此的研究设计将无法提供任何有用的信息。而要使因果推论具有不确定性，解释变量和被解释变量的取值必须由具体研究过程决定。

让我们举一个例子，以便读者更好地理解上述内容。对于国际冲突的观察，促使研究者推测政府组织形式是不是其与其他国家关系变化的决定因素。我们首先用一种探索的方式着手研究：仔细观察那些战事频繁的双边关系及一直和平的双边关系。我们发现，对于那些频繁发生战争的交战方来说，至少其中一方是现代独裁国家。此时研究

者已经获得一个比直观感受更加确切的研究假设了。当然,到目前为止研究者还没有一个非常清晰的整体认识,因此我们还不能说假设已经得到证实,下一步就可以根据解释变量来取样去证明假设。此外,研究者也需要控制其他决定军事冲突程度的因素,在控制它们之后再去考察政权类型和冲突之间的关系。

4.4.3 同时根据解释变量和被解释变量选取观察值

有目的地根据解释变量和被变量变量选取观察值是很危险的,这种做法稍不留神就会导致推论偏差。假设在选取的样本中解释变量和被解释变量同时变化且变化方向也符合假设,这就会导致最严重的推论错误。比如,假设我们试图研究威权统治(该政体往往会压制工会组织及劳动者需求)是否会推动高速经济增长。如果研究者选取的样本是:所有被选取的威权国家都有较高的经济增长率,同时所有被选取的非威权国家经济增长率都偏低,这样的研究设计就毫无意义,其结论也不能说明任何问题。因为选取的数据并没有代表性,因而无法告诉我们那些允许工会组织的民主国家的经济增长情况。

尽管根据解释变量和被解释变量选择观察值存在一定风险,但在样本获得受限的研究中,研究者还是可以在综合考虑解释变量和被解释变量的基础上选择该方法。但在操作过程中,研究者需要格外谨慎。举个例子,假设被解释变量的分布严重倾斜:大部分观察值都取一个值。此时如果还是根据解释变量选取观察值而不对被解释变量设任何限制的话,后者的取值完全有可能没有任何变化。在这种情况下,将不会从数据中获得任何有用的结果。事实上,如果无论解释变量怎么变化,被解释变量都不变的话,这正意味着两者之间没有因果效应。当然还有一种情况存在,就是此时因果效应很小但还不至于为 0。这种情况处理起来比较棘手。小样本研究中很难准确区分出无效应和微弱效应。对于这个问题,最直接的解决方法是增加观察值。另一种可行的策略是从解释变量中选取那些极端值,这样即使是微弱的因果效应也可能被发现。如果这两个方法都不能解决问题的话,研

究者只能尝试按照解释变量和被解释变量(但两者不能同时选择)去选择观察值了,这也可以增加发现因果效应的概率(其他相关建议见本书6.3节的内容)。

研究者可以运用一些抽样技术先按照解释变量来取样,同时也充分考虑到那些被解释变量中取值较少的观察值。这样一来,研究者就不能预先决定那些和被解释变量值相关的解释变量了。此外,在使用这种方法时,还要时刻警惕由于观察值受限导致的推论偏差问题。换句话说,根据解释变量和被解释变量选取观察值只适用于极少数的情况下,但即使这样,研究者也不能同时根据两者取样。①

为了便于理解,让我们举一个例子。假设某个国际组织中的某种联合成员资格可以有效抑制国家间的冲突。在这种情况下,我们的建议是只根据解释变量来选择观察值。依据这种方法,其数据组织结构是两个国家的配对组成一个观察值,同时这两个国家加入该组织的时间不同。如果我们很难确定这种特定会员模式是否存在的话,可用的观察值数量就会非常有限:可能最多只有20来对符合条件的国家组合。如果这些配对国家间很少爆发冲突,那么我们建议的方法可能并不可行。比方说,1 000对国家中只有一对爆发过战争。在这种情况下,这些样本在解释变量(特定会员资格)上存在差异,但在被解释变量上却没有任何变化。

如果研究者遇到这种情况,混合型的取样方法可能更为可行。所谓的混合型取样方法是指根据解释变量选取观察值(有些国家组合存在特定会员制模式,有些则没有),但比原计划选取更多观察值。然后将这些待用观察值再分成两类:在某个时间段曾经发生冲突的国家组合和那些没有发生冲突的组合。最终获得样本要在这两种组合中较多地选取那些发生冲突的观察值。该过程在操作过程中要独立于研究者对解释变量的既有认识。比如,可以选择所有发生过冲突的国家,然后再随机选择没有冲突的国家。如果组织会员资格和军事冲突之间存在

① 换一种说法就是,如果根据被解释变量和解释变量的边际分布去选择样本,通过该研究,研究者仍然可以了解其联合分布。

显著的相关关系，那么研究者就可以做出初步的因果关系。

阿图尔·科利(Atul Kohli, 1987)将印度作为研究对象，研究了国家在消除贫困中起的作用。他的研究就告诉了大家观察值有限的小样本如何影响因果关系的有效性及相应的克服方法。科利感兴趣的问题是发展中国家的政府结构及政权类型在减轻贫困中的作用。他所持的基本观点是：如果国家有明确的承诺要消除贫困、禁止社会上层阶级参与权力，同时拥有较强的组织能力的话，那么该国就能制定出有效的政策并获得预期目的。与此相反，那些没有明确的承诺、政权基础比较广泛及缺乏强有力组织的国家一般无法制定和推行行之有效的政策，即使它们公开宣称过要这么做。

印度被选为科利的研究对象，一方面是由于该国是他的研究兴趣所在，同时也由于他具备语言优势。他试图从印度的各州开始他的研究，“印度的联邦制特点决定了我们只能采用分解式的比较研究。在联邦政府领导下，印度各邦(或省)政府在制定和推行农业政策时发挥了显著的作用，各地在治理上存在的差异导致了减贫政策的不同效果”(Kohli, 1987:3—4)。在研究中，科利恰当地使用了“相同效应”假设。我们说过，该假设在强度上要弱于单位同质性假设。“相同效应”假设将解释变量的因果效应在不同层级地区的大小视作相同。也就是说，他认为意识形态、阶级基础及组织能力都是有助于扶贫的。这样的话就可以通过比较各个地区的被解释变量并施加“相同效应”假设来检验其因果假设。

使用印度各邦的证据非常具有优势。在科利看来，“较之其他独立国家，印度各邦的情况都比较接近”(Kohli, 1987:4)。接下来，科利凭借他两次在印度长时间的田野调查积累的认识并考虑到研究的其他限制，他选择了三个邦作为研究对象。需要指出的是，在这种情况下通过随机取样方法选取这三个邦是不明智的，因为随机取样只有在大样本研究中才起作用。研究发现，印度大部分地区的地方政府所采取的政权组织形式都不利于减贫政策的推行，因此这些地方政府很少有针对该问题的政策。被纳入科利研究的三个邦分别是西孟加拉邦、北方邦和卡纳塔卡邦。其中西孟加拉邦是印度各邦中唯一实行扶贫政策的地

方政府,这也是选取该邦的原因。北方邦推行的扶贫项目数量非常少,而卡纳塔卡的情况则介于上述两者之间。读者从中可以发现,这三个邦取样的根据是被解释变量:"三个邦在减贫政策上的努力组成了一个连续值,两端分别是最大值和最小值。"(Kohli, 1987:7)

当然研究面临的问题也是明显的:科利的取样过程实际上是同时根据解释变量和被解释变量进行的。在这种情况下,研究设计存在一定程度上的不确定性并且无法为因果推论提供有效的信息。也就是说,如果样本的选择是按照预先的假设来进行的话,由此针对假设所作的任何检验结果都是无效的。

读者也许会问,这项研究还有意义吗?如果科利只通过这三个邦的证据来检验假设的话,其结果的意义就不大。所幸的是,科利的工作远不止此。随着研究的开展,他发现之前的样本数量已经不足。而仅仅基于这三个邦的解释变量和被解释变量去支持假设的做法也没有被他采用,他选择进一步拓展样本。具体做法是通过引入邦的内部及其他国家的数据来增加观察值。以上这些小样本方法是下一章讨论的主要内容。此外,我们将在6.3.1小节中详细说明科利在这个小样本研究中所采用的策略。

从分析的总体水平上说,科利仍然可以进一步加强自己的因果推论。比如,他可以将印度所有地区的解释变量和被解释变量的值都收集到,然后在他的著作中另起一章简要描述各个邦的情况。这样做可以从总体上增加因果假设的真实性,同时有助于研究者采用一个更加系统的取样标准进行样本选择。

4.4.4 选择观察值使得关键解释变量为常数

社会科学家有时候在研究设计中会将作为选择标准的解释变量设为常量。这种做法存在明显缺陷:如果解释变量是常数的话,我们无法评估它的因果效应。因此,那些试图检验某些常数变量因果效应的研究用处一般不大。但是,任何研究都是某个文献或者研究传统的一部分(见本书1.2.1小节),同时研究者对即将从事的研究也有一定的经验

积累。凭借这些准备,在解释变量取值给定的情况下,研究者一般能够知道被解释变量的取值范围。因此,对于解释变量的另外一个取值,研究者完全可以大概获得相应被解释变量的取值并成为最先获得其因果效应的学者。

在解释变量为某一取值时,研究者首先需要对其作用做了一个非常合理的暂时性假设。随着研究的开展,解释变量的赋值产生变化,此前的假设就会被推翻。举个例子,在早期工业化作用的研究中,英克尔斯(Inkeles)和罗西(Rossi)(1956)从不同职业在社会声望上的差别这一问题出发对几个工业化国家进行比较。他们发现,这些国家虽然在很多方面存在差异(除了工业化之外),但也有很多相同之处。他们认为工业化是导致声望差别结构的重要因素。由于所有的样本都是工业化国家,因为作为解释变量其实是保持不变的,由此获得的因果推论并不可靠。当然,作为暂时性的结论是可以的。有趣的是,另外一批研究者在还没有实现工业化的菲律宾和印度尼西亚复制了上述研究,结果出人意料:在这些国家同样存在和那些工业化国家相似的职业声望分布。该结果就使得我们有理由怀疑工业化和社会声望之间的因果关系了(Zelditch, 1971)。

上述例子告诉读者,在早先研究计划中的解释变量没有变化时,后续研究如何克服该问题以获得有效因果推论。戴维·莱廷(David Laitin, 1986)的研究则告诉我们,同一个研究者也是可以在后续研究中克服这个问题的。莱廷研究了宗教信仰变化对尼日利亚的约鲁巴地区当地政治的影响。作者在文中讨论了自己之前基于索马里的研究无法解决该问题的原因。他指出,在索马里的研究中,作为解释变量的宗教是没有变化的。这就造成宗教和其他变量之间由于存在多重共线性问题(见 4.1 节的讨论)使得其因果效应无法被区分出来。“在索马里的田野研究中,我的研究问题是:宗教信仰变化是否会对政治产生独立影响?但随着研究的开展,我发现无法对该问题展开系统研究:你怎么可能在一个纯粹的伊斯兰世界里检验伊斯兰教对社会的影响呢?所有的当地人都说索马里语、继承了游牧民族的传统并沐浴在诗歌文化的光辉之中。因此,当地人行为上的共同点可能因为索马里的诗歌、游牧

及语言传统,而不一定是宗教传统。"(1986:186)为了解决这个问题,莱廷把他的研究对象转向尼日利亚的约鲁巴人,当地人由伊斯兰教信徒和基督徒构成,对于莱廷采用的具体研究方法,我们将在本书第5章详细讨论。

4.4.5 选择观察值使得被解释变量成为常数

如果样本中被解释变量的取值没有变化,我们也无法从中获得任何关于因果效应的有用信息。不过研究者如果能从文献中获得足够信息的话,还是可以获得有效因果推论的。

因此如果想知道为什么某一可能发生的结果却没有发生的话,研究者使用样本中的解释变量和被解释变量的取值就必须有变化。举个例子,我们对南卡罗来纳州的农场主没有使用化肥以保持土壤肥沃感兴趣。如果调查仅限于该州的话,我们很难得到有效结论,因为该地区所有的农场主行为习惯都很相似。被解释变量在这个例子中就没有变化。造成这种结果的原因是研究者没有引入新的信息。如果弗吉尼亚州的一些农场使用了化肥的话,研究者就可以把这两个州作为研究对象。这两个州之间的差异都可以作为潜在的被解释变量来解释其在化肥使用方面的不同。另一方面,如果之前所有的相关研究都在那些不用化肥的州进行的话,那么将那些使用化肥的州纳入研究范围本身就弥补了现有文献的不足。同时,新的样本也增加了检验因果假设的可能性。

再举一个例子,对于核战争的风险,政治学家有过很多耸人听闻的预测,这些预测也让整整一代人提心吊胆,但核战争却始终没有发生。尽管如此,对于核战争爆发的条件依然是很多学者的兴趣所在。该问题就是一个典型的被解释变量没有变化的例子。很多学者认为核武器没有被使用是因为其关键解释变量(世界上至少有两个核大国)的取值一直没有变化。但基于解释变量和被解释变量都不变的样本来估计因果效应是根本不可行的,除非研究者重新设计研究计划。在本书6.3.3小节中,我们将告诉读者解决这个问题的方法。

医学研究者有时也会采用回溯方法来检验因果假设。例如,疾病控制中心(CDC)一般会根据那些极端且不变的被解释变量来选择样本。CDC 首先选定一个"癌症组"——在同一地区患有同一种癌症的病人群体。然后在此基础上再去寻找可能致癌(被解释变量)的化学或其他因素(关键解释变量)。这种按照被解释变量极端值来选择样本的方法是很有效的,因为很多解释变量的取值都处在正常水平。CDC 的研究结果大都是负相关或者不确定的,但该组织也确实发现了一些可能的致癌化学因素。如果现有研究没有证据表明这些因素致癌的话,CDC 就会围绕这些因素开展深入研究。其具体采用的方法是按照该解释变量(该化学物是否存在)去选择样本以期更有把握发现因果效应。

医学采用的这种方法也被社会科学家所借鉴。研究者有时会关注一些特定的"政治群体":长期持有政治极端主义、鼓吹暴力的社区或地区,并试图寻找哪些"特殊"因素导致了上述特征。正如 CDC 所采用的方法,先用初步的研究表明它们之间可能存在某种相关。此时学者不应该直接由此获得结论,而是围绕它展开进一步研究,按照推测的解释变量去选择观察值,同时也让被解释变量——政治激进主义或暴力——的取值有变化。

4.5 结论

本章讨论了如何选择观察值以获得明确的研究设计从而消除推论偏差问题。当然,完美的研究设计是不存在的,研究者需要把对取样过程的批评和那些不完美但是很有用的对应策略结合起来,以解决研究中出现的问题。研究者最终凭借的研究设计应当具备以下条件:依靠理论并根据解释变量选择观察值的同时,也让被解释变量有变化。除此之外,根据被解释变量选择样本在一定情况下也是有用的,但是需要非常谨慎地操作。总的来说,研究设计需要达到的首要目标是尽可能多地获取检验理论的相关信息,同时避免引入可能损害推论质量的偏差。

5 在研究中需要避免的问题

在本书第 4 章中，我们讨论了在研究设计中如何通过适当的取样方法来获得推论。但仅完成样本的选择并不意味着就能获得有效的推论，这仅是研究开展的必要前提，后期出现的一些分析上的偏差依然可以完全毁掉前期的准备。而本章的内容就是要讨论在取样之后如何去发现那些导致无效推论及估计偏差的原因，并将其纳入可控范围内。我们将讨论在研究中如何提防这些问题以便有效应对。

在讨论无效推论和估计偏差之前，先让我们回顾一下本书 2.7 节及 3.4 节中关于如何评价推论的内容。在研究设计之后，研究者的注意力应当集中到两个关键问题上：无效性和偏差。在本章中我们将详细讨论这两个推论标准。为了理解这两个概念，读者可以将任何一个推论视作对位于某一区间内的某一特定值的估计值。比如，我们猜测一个人的年龄是 40 岁，上下偏差 2 岁。那么 40 岁就是最佳估计值，该估计值位于 38 岁到 42 岁这个区间内。区间的大小表示不确定的程度。我们都希望通过选定一个区间，使得估计值在大多数情况下位于这个区间内。因此，无偏性指的是区间中心恰好是正确的估计值，而有效性是指缩小该区间。

这两个概念也适用于本书之前章节所讨论的描述及因果推论。打个比方，在估计教育对收入的影响（即每增加一年的教育所带来的额外收入）时，研究者将获得一个点估计值和一段围绕该值的区间用以表示对该估计结果的不确定性。这个区间越窄，意味着估计的有效性越高。同时，如果该区间能向准确估计值集中，就意味着估计是无偏的。当然，我们也希望区间的上下限能准确反映出估计值的不确定性。

本章首先讨论如何在研究的初始阶段提高信息使用质量，其中将重点关注导致偏差和无效的四个原因。5.1 节将讨论测量误差，该问题会导致估计结果产生偏差并降低其有效性；5.2 节讨论由于遗漏变量所导致的偏差；5.3 节的讨论重点放在控制上：控制那些可能降低估计有效性的无关变量；在 5.4 节中我们将讨论内生性问题：被解释变量反过来影响解释变量。最后，5.5 及 5.6 节将分别讨论对解释变量的随机赋值和非实验控制方法。

5.1 测量误差

研究者选好样本后，接下来的任务就是对变量进行测量。在社会科学中，针对任何观察值的测量都不可能百分之百精确，因此在研究中测量误差是不可避免的。

很多社会科学研究都试图对误差的大小进行估计并尽量减少其带来的影响。定量研究使用的测量工具更为精确（或更加量化），但这并不必然意味着测量结果更为准确。一般来讲，在追求信度（reliability）时往往以牺牲效度（validity）为代价。前者是指用不同测量方法对同一现象进行测量以期获得相同的结果，而后者则是指测量结果必须能够反映出研究者想要的测量内容。相比之下，定性研究也会试图获得准确的测量，但一般来说，其测量准确性不如定量研究。

定量测量和定性观察在本质上是相似的。定性研究者一般用文字来表示类别，而定量研究则多用数值表示。这两种研究方法都会用到定类测量（nominal measurement）、定序测量（ordinal measurement）和定距测量（interval measurement）。在定类测量中，研究者假设观察对象的不同类别间没有特定顺序。这些类别一般按照法律或者制度设置区分。例如，比较政治学可能将不同国家的政治制度分为总统制、议会制和威权制度。定序测量将研究对象按照特定次序分类。比如，定性研究者按照工业化程度或者军队规模将国家分成三个或四个类别。定

距测量往往使用连续变量，比如跨国交易数额。

定量及定性研究所使用的测量方法不在于其基于的理论，而是数据所呈现出来的形式。在定性研究中，研究者多使用“更多”或“更少”、“更大”或“更小”、“强”或“弱”这样的词汇。但是在定量研究中，研究者则直接使用数字。

让我们举一个例子，国际关系领域中的研究者大都认为“战争中死亡人数”并不是测量战争对世界格局影响的好指标。权力均衡理论认为，相较战争的严重程度，主要参战方在战后的变化更能体现研究者所关注的“不稳定性”（见 Gulick，1967；Waltz，1979:162）。然而，在实际研究中，很多定性学者为了追求效度，往往因为测量偏差问题导致信度的丧失。打个比方，如果缺乏准确的定义，究竟什么才算“随之发生”呢？事实上，不同说法其实暗示着研究者对于变化的判断是基于系统结果作出的，而这种结果往往会导致对假设中因果方向的估计存在偏差。

如何在一个不太能有效反映我们感兴趣概念的定量指标和一个不太精确且可能带有偏差的定性指标间作出取舍，对此并没有现成的标准可以参照。但不管采用哪个指标，研究者都应当说明推论中存在的不确定性。采用定量方法的研究者应当提供测量值的标准误，而定性研究者则应当通过较为谨慎的措辞说明估计中存在的不确定性。这两种方法之间的差别仅在于用不同的表述方式去表达相同的研究想法。

此外，这两种方法还有一个相似之处：类别或者测量通常是人为设置而非固有的。比如，将国家划分为民主/威权或者总统/议会制度；根据工业化水平对国家进行排序，等等。这些分类或者测量都取决于具体研究的需要。

显而易见，上述提到的那些分类或测量方式都没有普遍标准，最终采用的测量方法取决于研究者试图回答的问题。分类标准越接近研究者最初理论及实证设想越好。这样的标准再次说明了如何分类是研究者有意设计的结果。采用比例代表制作为主要代议制度的国家数量取决于研究者对“议会制”的分类及比例代表制度的定义。在国际关系领域中针对跨国交易的研究是否采用连续变量测量交易金额取决于单笔

交易的构成及国境线的定义。类似地，民主党候选人在国会选区中获得的选票比例同样取决于研究者的分类：全国 435 个国会选区中对两党的定义都要相同。

在本节中，我们使用不同的测量方式（定类、定序和定距）也依赖于测量的理论目的。举个例子，在政治学文献中，种族一般被视作较为典型的定类变量。在美国，该变量一般有黑人、白人、拉丁裔、印第安人和亚裔五类。但是具体到各个种族内部，其成员对于所属种族的认同感却不尽相同。因此，研究者可以按照该认同感进一步对种族进行分类，例如那些有强烈认同感的成员比重。或者，如果研究者对族群大小感兴趣的话，也可以用定距变量来区分。总之，关键在于选择符合研究目的的测量方法。

如果测量没有理论依据的话就可能会出现问题。比如，研究者往往根据年龄大小将研究对象分为青年、中年和老年三类。之所以做出这样的区分，是因为年龄是随自然变化的且容易测量。一般来说，这样的分类已经足够了。但对于某些研究，这样的区分显然不够精确。在分类过程中会犯所谓的“分类错误”（grouping error），该问题可能会对整个研究结果造成很大的影响。因此研究者应当谨慎以避免此类错误。我们认为避免“分类错误”的一个原则是不要因为有些数据看起来没用就丢弃它们。

有时候，研究者也会犯截然相反的错误：将不连续变量赋成连续或者定距变量。定距变量并不一定就比定序或者定类变量好。例如，研究者试图通过问卷对宗教信仰及参加宗教活动的频度进行测量。如果问卷设计得当，参加宗教的频度可以被设定为定序甚至是定距变量，这往往取决于选取的测量工具。但是如果将宗教信仰也设定为定序变量并按照数值来排列就没有多大意义。在后一种情况下，定序或者定距变量要么不存在，要么会造成测量误差。

在定类、定序和定距测量方法间的选择取决于表述的丰富性及比较的便利性。让我们以国际组织中的投票规则为例，投票规则之所以重要，是因为在很大程度上它反映了各个成员国的主权，同时也因为它可以决定决议是否通过、资源如何分配及对服从该组织委托授权的

期望。

对于国际组织中的投票规则，研究者可以将其分为一票否决制（例如国际联盟理事会盟约第 15 条规定采取一票否决制）、特定成员否决制（在联合国安理会的决议中，如果五个常任理事国中任何一国对议案投否决票，该议案就不能被通过）、绝对多数制（欧盟在进行内部市场决议时经常使用该规则）以及相对多数制（联合国大会经常采用该规则）。显而易见，上述每种投票规则都会赋予其成员国不同的交涉动力。如果研究者仅仅对其中某个规则感兴趣（比如一票否决制），那么如何定义分类就尤为重要，也就会避免将其他规则纳入研究范围。此时，定类指标相对来说就更加合适。

当然，研究者也可以按照严格程度对上述规则进行排序：从最严格的（全体通过）到最宽松的（相对多数）。如果试图回答的问题是规则严格程度对成员间交涉模式或者特定投票结果的分布规律影响的话，定序分类就显得非常有必要了。不过应当提醒读者注意，在这种情况下，特定成员否决制和绝对多数制这两种规则就很难排序，因为它们内部都有一系列不同的安排。例如，特定成员否决制既可以是某一个成员拥有否决权（这种情况下该规则接近独裁制度），也可以是除了少数成员外的全体成员否决制（欧盟和国际货币基金组织目前采取的规则与此类似）。前者旨在防止任意两个国家在涉及内部市场等议题上投否决票。对于国际货币基金组织来说，和前者不同的是其名义使用绝对多数制，但同时赋予美国和欧盟否决权。因此，后者采用的投票规则既可以是绝对多数制，也可以划为特定成员制。

基于上述描述，如果研究者试图用定距变量区分这些投票规则的话，可以使用通过决议所要求的国家在某个指标上所占的比例（或资源比例：依据国民生产总值、对组织的贡献；或成员国的人数）来测量规则通过的严格程度。

显而易见，不同测量方法会产生不同的测量结果。例如，虽然人口和国民生产总值都能测量资源禀赋，但产生的结果却不同。因此，不同测量方法在精确性上的优势可能会被方法选择的任意性及合并不同结果的复杂性所抵消。读者应当认识到，不同方法都有各自的优势和局

限性，研究者应当根据研究目的作出取舍选择。

在以下两个小节中，我们将讨论定性研究中由于测量误差导致的后果，得到的某些结论可能会出乎读者的意料。大多数人都同意如果对某一变量的测量一直偏高，那么这种系统性测量偏误会导致估计结果偏差。同时，这种偏差并不会随着存在误差的观察值数量增加而消失。当然并不是所有类型的系统性测量误差都会使得推论产生偏差。和系统性测量误差相比，非系统性测量误差对结果的影响就不那么容易被发现了。以下内容将从两个方面对非系统性测量误差类型进行讨论：被解释变量中存在误差和解释变量中存在误差。一般来说，如果被解释变量存在非系统测量误差，就会导致估计无效进而获得错误结论，同时研究者也很难发现问题所在。换句话说，被解释变量中的非系统性测量误差不会带来估计偏差，但是会显著降低估计效率。对于解释变量，如果存在非系统性测量误差的话，会使得推论偏向研究者的预期。因此，了解这些测量误差的特性能让我们对研究作出改进从而尽可能避免它们。

5.1.1 系统性测量误差

我们将在本小节讨论系统性测量误差可能导致的后果。例如，对某个变量的测量一直偏高，那么这样的系统性测量误差会使得因果估计偏差和不一致。以下的任务就是弄清楚不同的系统性测量误差会导致什么样的偏差。无论是定量研究还是定性研究，对数据的选择有可能导致误差：研究者往往选择那些符合自己预期的数据。在定量研究中，研究者可能会使用带有测量误差的数据，因为该数据是唯一可用的；而在定性研究中，该误差可能由研究者的主观判断造成：研究者在进行研究之前往往已经有了假设，并希望证明其正确性。

显而易见，任何系统性测量误差都会导致描述性推论出现偏差。①

① 此时也有一种例外情况：正的系统性误差与负的系统性误差相互抵消。但这种情况非常少见且可被视作非系统性测量误差。

让我们看一个简单的例子:研究者高估了受访者的年收入——比实际收入多1 000美元。那么基于该数据计算出来的平均年收入就是多出相同的数额。如果研究者试图检验大学教育对年收入的影响,刚才的测量误差就不会影响其因果推论。再假设接受调查的大学生群体平均年收入为3万美元,而作为对照组的没有大学学历的受访者的平均年收入是2.5万美元。由此推断大学教育可以增加年薪0.5万美元。如果两个组的平均年薪都被高估相同的数额(比方说,都多1 000美元),那么大学教育对于收入作用的估计值仍然是0.5万美元(3.1万美元和2.6万美元之间的差额)。由此我们可以得出如下结论:如果所有样本中的系统性测量误差都相同的话,该误差不会造成因果推论出现偏差(读者可以回顾一下3.3.1小节中关于单位同质性假设的另外一个版本:同效应假设的讨论就是明白这一点)。

然而,如果只有一组样本出现了系统性测量误差,那么此时情况就不同了。为了在访问者面前显示自己的能力,假设那些大学学历的受访者都报高了自己的收入,但作为对照组的那些受访者汇报的是真实收入。在这种情况下,对教育对收入作用的描述性和因果推断都会产生偏差。如果研究者能预料到有些受访者可能故意报高收入的话,就会在问卷设计上多下功夫或从其他途径获取更准确的信息。即使发现该问题时数据收集工作已经完成,同时也没有机会获得其他信息,研究者至少能确定偏差的方向,并在后续阶段进行相应的修正。

为了让读者更好地理解这一点,让我们再看一个国际关系领域中关于区域一体化的例子。和该领域其他相关研究不同,针对该议题的研究有时会采用定量指标去检验假设。不过并不是该领域的所有概念都适合这样做。例如,主权国家在多大程度上将决策权力转移至国际组织,这个概念在这个领域中非常重要却很难定量测定。因此,研究者大都是用定性的测量方法,其划分标准则基于研究者对该问题的了解程度(Lindberg & Sheingold, 1970:71,表3.1)。已有研究曾使用的分类包括“精英价值观互补”(elite value complementarity)或“决策风格”(decision-making style)(见Nye, 1971;Lindberg & Sheingold, 1971)等较为主观的指标,并试图通过这种分类去检验变量间的因果关系。

当然，上述方法事实上是出于对测量有效性的考虑：研究者对信息进行编码并检验它们是否与所要测量的概念相关。但是该方法可能引起主观测量误差。在研究过程中，研究者必须严格要求自己并始终做到约束个人偏好，在对解释变量进行编码时避免带进自己的理论预期。作为读者来说，也要判断研究者在多大程度上避免了该问题。

对于这种情况，我们的建议是：第一，尝试借鉴其他研究者在与该问题完全不相关的研究中所持的判断。不管是定量研究还是定性研究，测量方式都有一定程度上的随意性。该建议可以使得研究者不会被还没成型的研究假设所影响。该建议在定量研究中使用得较为普遍，该领域的研究者经常在研究中借鉴和使用其他研究使用的测量方法。其实该策略也同样适用于定性研究。举个例子，利用不同研究者对同一现象的不同解释可以建构一个共享编码的数据库，也可以基于研究者的经验和讨论建立一个含有标准类别的定性数据库。这些工作都有助于后续学者检验他们的研究假设。此外，如果你首次使用某个变量，请一个对于研究并不了解的其他领域的专家帮助你编码是一个很有用的方法。你可以将你的田野调查笔记和访谈录音给他，然后让他基于这些信息进行编码，再比较他的编码结果是否与你的相同。这种重复编码工作可以增加使用定性变量的信心。总之，找一些经验丰富的研究者去复核你的测量工作，对研究的提升作用非常明显。

5.1.2 非系统测量误差

不管是定量研究还是定性研究，非系统性测量误差都是研究者必须面对的一大难题。[①]非系统性误差不会导致变量的测量产生误差。在本节中，我们定义的非系统测量误差是那些有时导致测量结果偏高有时导致偏低，但平均起来是准确的误差。随机性误差经常导致统计推论的无效，但是不会使描述性推断出现偏差。这一点已经在2.7.1小

① 这是因为我们的能力尚不足以准确测量世界的种种现象，还是因为真实世界的随机性？这就成了一个难以琢磨的哲学问题，可能有很多种不同的答案（见本书2.6节）。当然，无论研究者接受哪种立场，结果都是一样的。

节中详细讨论过了，因此此处我们将不再讨论随机性误差对描述性推断产生的影响。本小节的重点是讨论这种误差对因果推断的影响。

在估计因果效应时，解释变量和被解释变量中的随机测量误差会导致不同的后果。如果被解释变量有随机性误差，会降低因果推断的有效性，但并不会使推论出现偏差。这种误差可能导致研究者高估或低估因果效应。但总体而言，估计值是准确的。事实上，被解释变量的随机测量误差和现实世界中普遍存在的随机误差没有什么区别，甚至很难区分开来。

如果解释变量出现随机性测量偏差，同样会导致因果推断的无效性并高估或低估估计值。但和被解释变量中存在随机性误差不同，解释变量中存在随机误差会导致因果效应的估计值总是低于真实值。也就是说，如果被解释变量和解释变量之间确实存在因果关系，解释变量中的随机性误差会导致结果错误地显示两者之间没有或者只有很弱的关系。相反，如果被解释变量和解释变量之间实际是负相关的，解释变量存在的随机性误差就会导致因果关系趋近零。

虽然被解释变量和解释变量中的随机性误差对研究产生的影响不小，但是研究者一般很难觉察到这种不同。为了更好地解释这两种误差的影响，我们将会在之后的章节中提供相应的公式证明，并辅以图表和例子。我们先从被解释变量中的随机性测量误差开始。

5.1.2.1 被解释变量中的非系统性测量误差

非系统性测量误差也被称为“随机测量误差”。虽然该误差存在于被解释变量中不会导致因果推断产生偏差，但会降低估计的有效性。这种无效性都会导致因果效应被高估或低估。因此，被解释变量的测量误差会增加推论的不确定性。从这个意义上来说，被解释变量存在随机测量误差所导致的问题和利用小样本作推论所导致的问题有相似之处。在这两种情况下，研究者都会面临信息不够的困境。综上所述，被解释变量的随机测量误差会导致因果推论有效性下降并增加推论的不确定性。

如果研究者同时使用多套数据，被解释变量中的随机性误差会使

估计结果不稳定:其中一些数据可以获得显著的正向因果关系,但基于另一些数据的估计结果却发现因果关系根本不存在甚至负相关。要知道真实因果关系只可能是这三种结果中的一种。在这种情况下,研究者就很难通过这些数据获得系统的描述及因果推论。这样一来,我们就将面临信息量不足的问题,无法找到实际存在的因果效应,因为它往往被被解释变量中的随机性误差干扰(这种误差也可能导致推断不确定性增加)。尽管这个问题比较严重,但不管定量还是定性研究者都没有很有效的处理方法。唯一能做的就是提高现有数据的测量精度或者使用其他没有测量误差的数据。如果能用具体数值表示出造成的后果,研究者就能提高研究结论的质量。明确掌握研究结论的不确定性能够促进研究者开展后续研究,在进一步的研究中提高被解释变量的测量精度或者进一步增加观察值的数量。此外,对不确定性的掌握在制定研究设计时也很有帮助,研究者据此可以在提高观察值测量精度和增加观察值数量之间做出权衡。总之,研究者的最终目标是尽可能地获取和假设相关的信息:既可以通过从现有样本中挖掘新信息,也可以引进更多的新数据。

现在让我们举一个被解释变量中存在随机性误差的例子。假设研究者试图研究经济发展对暴力犯罪的影响,其研究样本基于发展中国家或某个发展中国家中的不同地区。研究者首先需要测量一段时间内不同地区的暴力犯罪数量,也就是该研究的被解释变量。在实际情况中极有可能出现这种情况:该变量可能存在测量误差问题。虽然从整体上通过测量获取的犯罪水平和真实水平接近,但有些地区的数值可能被高估,而在另外一些地区却低估了真实水平。

研究者用失业率来测量该因果假设的解释变量:经济发展水平。一般认为官方的失业率数据较为可靠,因此测量误差并不严重。这样,该因果假设就具体到失业率对暴力犯罪的影响上了。对于该因果假设的估计结果,研究者应当明了其存在严重的不确定性,也就是说不同的估计工作获得的结论可能完全不同。这种不确定性来自该研究的被解释变量——犯罪率——存在测量误差。不过即使是一般读者,也会相信这两个变量之间确实存在很强的因果关系。据此,研

究者调整了研究方案:设置更多的观察点并延长观察时间。即使这样,测量误差依然存在且观察值数量仍然不足,估计结论依然不确定。一般而言,我们有两个途径提高由于结论不确定造成的估计低效问题:第一,提高现有样本的测量精度。就这个例子来说可以参考警方的犯罪记录来减少测量误差。第二,增加样本数量,但要保证新增加样本的测量误差较小。上述两种途径均能够增加信息量进而降低因果推论过程中的不确定性或低效问题。上文的例子也反映出信息量比单纯的样本数量更重要。

为什么这么说呢? 让我们先通过下面的图来简单说明,然后再给出公式化证明。图 5.1 的横轴表示失业率,纵轴则表示暴力犯罪。首先假设图中的失业率 4%和 7%反映的是实际失业水平。

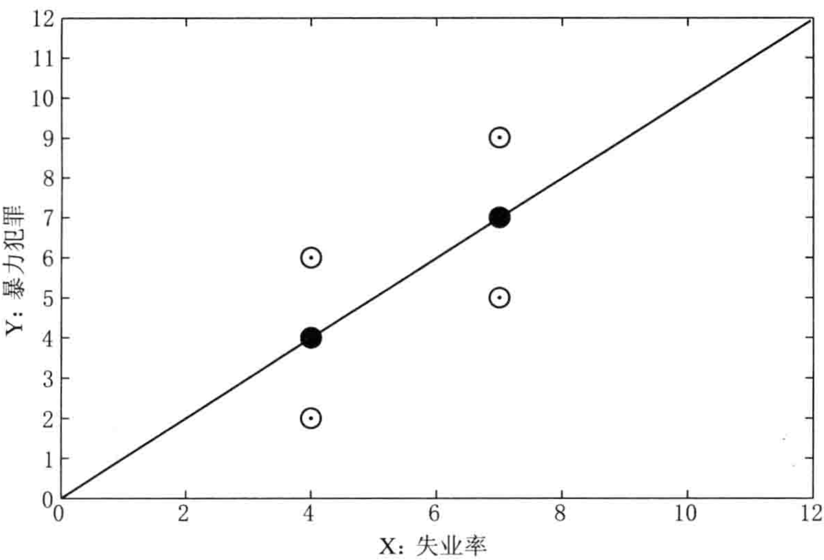


图 5.1 被解释变量中存在测量偏误

图中的这两个实心圆可以看作所有变量都没有测量误差的情况。出现这种情况有两种可能:第一,基于大样本数据,同时所有的观察值都恰好落在这两个点上;第二,只有两个观察值,但它们都不存在系统性测量误差。然而在真正的研究中,出现这两种情况的可能性非常小。因此,从图 5.1 中,我们可知对于那些更具一般性及更加复杂的研究对

象来说，被解释变量出现测量误差将会导致的严重后果。如果读者留意该图中那条拟合线是怎样生成的，便可以理解这句话的意思。

现在假设对暴力犯罪的测量存在非系统性误差。之前提到，在这种情况下，测量值在平均意义上和真实值相同。为了体现出这一点，让我们在图上标记四个空心圆，分别位于实心圆上下但保持相同距离。① 如果基于这六个点拟合估计直线的话，该拟合线依然是原来的拟合线。读者应当注意这条拟合线能够使得预期的误差最小化，也就是说四个空心圆到直线的垂直距离最短。

但是在这种情况下，新拟合线的确定性要差得多：如果我们稍微增加或减少一下线的斜率，新生成的线依然能很好地拟合图中的六个点。此外，即使斜率保持不变，拟合线在图 5.1 中也可能向上或者向下平行移动，进而无法对单个观察值进行精确的预测。这种无效估计的出现正是由于被解释变量中的测量误差所致。在这种情况下，估计结果是无偏的：如果估计很多次，平均起来估计结果就会趋近真实值。但对于其中某一次估计来说，获得的估计值可能和真实值相差甚远。

对 Y 中存在测量误差的公式化分析

在一个简单的一元线性模型中，首先假设被解释变量存在测量误差但解释变量没有误差。我们的目的是估计参数 β 的值：

$$E(Y^*) = \beta X$$

同时 Y^* 的方差是：

$$V(Y_i^*) = \sigma^2$$

下标 i 表示对所有观察值 $i = 1, \cdots, n$，Y^* 的方差都一样。②如果我们没有观察到 Y^*，而是观察到了 Y，那么它们之间的关系如下：

---

① 我们再次假设空心圆背后有大量的观测数据，只是这些数据刚好落在这四个点上，或假设几乎不存在随机变化。

② 熟悉统计学的读者可将其称为同方差或不变方差的性质。

$$Y = Y^* + U$$

该表达式意味着观察到的被解释变量 Y 等于真实值 Y^* 加上某个随机测量误差 U。为了表示 U 仅包括非系统性测量误差，需要进一步作如下假设：如果估计可以重复多次，平均来说误差可以相互抵消，即 $E(U)=0$；同时，误差和真实被解释变量 Y^* 不相关，即 $C(U, Y^*)=0$，和解释变量也不相关 $C(U, X)=0$。①此外，每个观察值 i 的测量误差的是方差 $V(U_i)=\tau^2$。其中，τ^2 越大就意味着 Y 的测量误差越大。τ^2 等于 0 则意味着不存在测量误差，此时 Y 等于 Y^*。

被解释变量中的随机测量误差如何影响对 β 的估计呢？使用一般的 OLS 估计方法但用 Y 代替 Y^* 时，β 的估计值如下：

$$b = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

然后进行多次估计并计算 β 的均值：

$$\begin{aligned} E(b) &= E\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \\ &= \frac{\sum_{i=1}^{n} X_i E(Y_i)}{\sum_{i=1}^{n} X_i^2} \\ &= \frac{\sum_{i=1}^{n} X_i E(Y_i + U)}{\sum_{i=1}^{n} X_i^2} \\ &= \frac{\sum_{i=1}^{n} X_i^2 \beta}{\sum_{i=1}^{n} X_i^2} \\ &= \beta \end{aligned}$$

上述计算过程可以告诉我们，即便被解释变量存在测量误差，OLS

① 这些关于误差的假设说明观察到的被解释变量期望值和被解释变量真实值的期望值相同：$E(Y) = E(Y^* + U) = E(Y^*) + E(U) = E(Y^*) = \beta X$。

估计结果依然是无偏的（就是说，平均来说 b 等于真实的因果关系 β）。该结果和公式(3.8)中（被解释变量没有测量误差）的计算结果相同。

除了需要说明估计结果是否有偏外，当被解释变量存在测量误差时还要对结果的有效性进行评估。我们依然采用最一般的 OLS 方差计算方法：

$$V(b)=V\left(\frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2}\right) \tag{5.1}$$

$$=\frac{1}{(\sum_{i=1}^{n}X_i^2)^2}\sum_{i=1}^{n}X_i^2V(Y_i^*+U)$$

$$=\frac{\sigma^2+\tau^2}{\sum_{i=1}^{n}X_i^2}$$

公式(5.1)清楚地显示出当被解释变量存在测量误差时，获得估计结果的有效性降低了——读者可以比较公式(5.1)和公式(3.9)。在这种情况下，有效性的损失取决于被解释变量测量误差 τ^2 的大小。

5.1.2.2 解释变量中的非系统性测量误差

在之前的内容中我们提到，解释变量如果存在非系统性测量误差会在描述性推论中造成的后果，和被解释变量出现非系统测量误差所造成的后果相似：测量结果时高时低，但平均来说测量结果和真实水平相同。此外，解释变量的随机误差会使得估计结果变得不确定和低效，这点也和被解释变量存在非系统误差导致的结果相同。除了这些相似之外，读者应该明白，解释变量中的随机误差引起的后果更加严重：会导致对估计结果产生系统性偏差。一般来说，这种偏差使得估计出来的作用趋近于零。换句话说，解释变量中的随机偏误使得它和被解释变量之间本来存在的关系被弱化以至于不容易被发现。如果研究者用几套数据来检验假设，一方面不同数据获得的结果之间存在很大差异，这种情况和被解释变量存在随机误差相似；另一方面，基于这些数据的估计结果都存在系统性偏差，导致估计结果低于真实作用。

即使研究者已经意识到该问题的存在并更加谨慎地分析存在问题的变量，上面提到的那些后果还是能被妥善处理的。碰到这种情况，我们一贯的建议就是提高对变量的测量精度，只有这样才能有效降低估计结果的低效和不一致问题。

让我们再回到那个欠发达国家中失业率对犯罪作用的例子。现在假定犯罪数据的质量得到改进：这次犯罪报告可以很容易地从相关政府部门获取且内容准确。但失业牵涉到政治问题，因此还是不能准确地测量出来。政府也不允许进行系统调查，研究者便决定通过直接观测的方式获取失业情况（就像之前例子中用到的直接观察法一样）：驾车穿过村子时去观察有多少看上去无所事事的人，并据此推算失业率水平。要注意，研究者在不同地点实施观察时的时间和天气情况都须一样，这些因素都会影响人是否出现在户外。这就意味着对失业水平的测量存在很多随机性的误差。当然如果研究者观察的样本足够大，这种误差可以被消除。但如果研究者仅通过两个村子去估计因果关系的话，这种误差就会导致估计结果低效：比如取样时间是周日，也许有很多人选择户外活动；而取样那天恰逢下雨的话，大多数人都会待在室内。如果研究者意识到这些问题并考虑到不同时间和不同天气情况的话，基于大样本的研究就会消除这些因素对失业水平估计的影响。这和上面例子中对暴力犯罪的测量工作遇到的情况类似，读者需要明白，解释变量存在误差会导致完全不同的后果。

图 5.2 展示了该问题存在时会导致的后果。实心圆表示被解释变量和解释变量都没有测量误差。①因此连接实心圆获得拟合线的斜率便是失业对犯罪因果效应的正确估计。为了清晰地表示出测量误差所造成的后果，我们在实心圆的左右两边分别放两个空心圆，它们代表解释变量的测量误差。此外，空心圆围绕相应的实心圆左右对称，这样就使得解释变量的测量总体来说是准确的，因为此时空心圆的均值恰好等于实心圆的取值。图中的虚线是对空心圆的拟合，由此导致和实线

① 我们继续假设每个点所代表的数据要么没有随机变化，要么就是很多点碰巧都落在这些点上。在本书 5.1 节中，该假设的目的是将叙述重点集中在问题本身上。

的区别正是来自解释变量的测量误差。需要提醒读者注意的是，拟合线的确定原则是在给定解释变量取值的情况下，使得被解释变量真实取值到拟合线的垂直距离最小。在这种情况下，我们就说用拟合线预测被解释变量时，和其真实取值之间的误差最小。

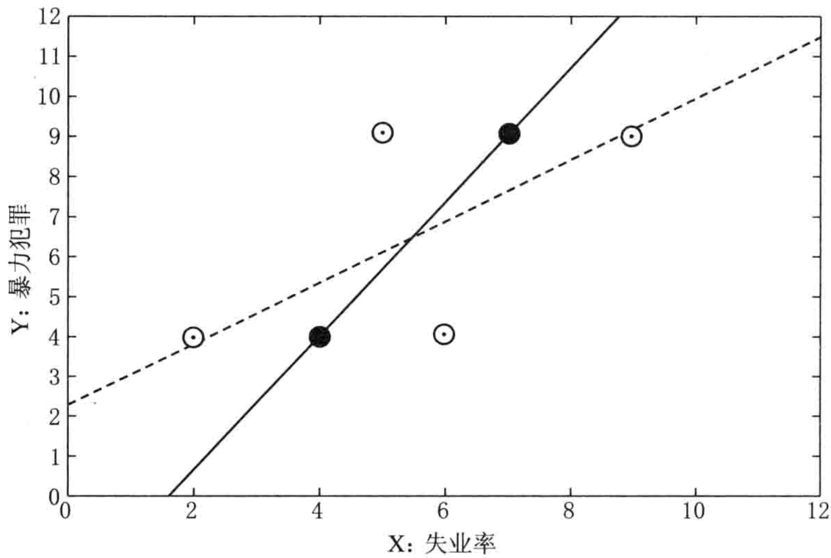


图 5.2　解释变量中存在测量偏误

此时对失业和犯罪之间关系的估计受到随机测量误差的影响，图 5.2 中拟合虚线的斜率明显小于实线的斜率，这意味着估计出来的效用小于真实效用。因此，读者凭借对解释变量测量误差的理解，就可以推断出失业对于犯罪的作用要比估计获得的要大。

以上讨论了解释变量存在测量误差所导致的后果，这些内容为读者提供了两个比较实用的指导原则。

1. 如果没有发现任何因果效应，同时也不清楚偏差的方向，那么真实的因果关系也就无从确定。此时，手头的数据无法告诉研究者真实的因果关系是正是负，抑或根本不存在因果关系。如果出现这种情况，研究者如实表述即可。

2. 研究者已经知道解释变量存在随机测量误差，如果研究发现微弱的正向作用，那么读者便可以利用本节的内容推测真实作用应当更

大;同样,如果发现微弱的负向效用,那么真实作用依然为负且更大。

在很多定量研究中,测量误差问题是不可避免的,因此以上指导原则可被广泛使用。

以下关于这些结论的简短公式化讨论能够让读者了解它们的适用性。第一,以下分析适用于只有一个解释变量的情况。对于多个解释变量,这些推论并非在任何情况下都适用。如果研究中有很多解释变量但只有一个存在严重的随机测量误差的话,这些分析就是适用的。不过,如果研究者有很多的解释变量且每个都有测量误差,此时试图同时估计其效应的话,以下公式就需要做一定的拓展以发现偏差所在。虽然在定性研究中也会有很多解释变量,但研究者通常是依次讨论它们的作用而非同时检验。正如本章 5.2 节的讨论,这种方式会带来诸如遗漏变量偏差问题。即使如此,在定性研究中,这种情况也会经常遇到。

对 X 存在随机测量误差的公式分析

首先定义模型如下:

$$E(Y)=\beta X^*$$

在研究中未能观察到解释变量的真实值 X^*,观察到的是 X,其中

$$X=X^*+U$$

U 是随机测量误差:$E(U)=0$,意味着均值为 0;$C(U, X^*)=0$,意味着和真实解释变量无关;$C(U, Y)=0$,意味着和被解释变量也无关。

如果研究者没有观察到 X^*,而是用带有测量误差的 X 来估计 β 的话,将会产生什么后果呢?这种情况即使在定性研究中也会碰到:存在测量误差但没有对研究结论作出相应的调整。下面的公式将演示该问题是如何给估计结果带来偏差的。基于的公式依然是公式(3.7),用到的数据则是观察到的 X 和 Y。

$$b = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \tag{5.2}$$
$$= \frac{\sum_{i=1}^{n} (X_i^* + U_i) Y_i}{\sum_{i=1}^{n} (X_i^* + U_i)^2}$$
$$= \frac{\sum_{i=1}^{n} X_i^* Y_i + \sum_{i=1}^{n} U_i Y_i}{\sum_{i=1}^{n} X_i^{*2} + \sum_{i=1}^{n} U_i^2 + (2\sum_{i=1}^{n} X_i^* U_i)}$$

公式(5.2)清楚地表明对 b 的 OLS 估计值会出现偏差，此时 $E(b) \neq \beta$。此外，之前假设 U 和 Y 及 X^* 都不相关，那么公式(5.2)最后一行中的 $\sum_{i=1}^{n} U_i Y_i$ 和 $\sum_{i=1}^{n} X_i^* U_i$ 都等于 0。因此该表达式可被简化为①：

$$b \approx \frac{\sum_{i=1}^{n} X_i^* Y_i}{\sum_{i=1}^{n} X_i^{*2} + \sum_{i=1}^{n} U_i^2}$$

相比公式(3.7)的 OLS 估计值表达式，β 估计值的分母部分多了 $\sum_{i=1}^{n} U_i^2$。这部分表示 X 中测量误差的样本方差。如果解释变量不存在测量误差，那么这项就等于 0，此时该表达式等于公式(3.7)的估计表达式。

但是在实际研究中，解释变量总是存在或多或少的测量误差，因此 $\sum_{i=1}^{n} U_i^2$ 总是为正数。又由于该正数出现在估计值的分母中，这导致 b 趋向为 0。换句话说，如果真实因果效应是较大的正数，解释变量的随机测量误差会使得研究者误以为 b 是一个较小的正数；相似地，如果真实因果效应是较大的负数，该问题会使研究者误以为 b 是一个较小的负数。

最后，上述公式同样可以演示在解释变量存在测量偏误的情况下，估计结果是低效的。不过一般而言，偏差是更严重的问题，因此我们在以下内容中先讨论偏差。

① 由于该表达式只能在大样本中成立，实际上这里是在分析一致性而非无偏性（本书 2.7.1 节）。更确切地说，当公式(5.2)中括号中的项是除以 n 时，该项会随着 n 趋向于无穷大而消失。

5.2 遗漏变量偏差

在估计某一变量对另外一个变量的作用时，大多数从事定性研究的社会科学家都认识到应该控制其他变量以防止由于遗漏变量问题而导致的虚假因果效应。具体操作方法包括：约翰·斯图亚特·穆勒(John Stuart Mill, 1843)提出的差异法和类同法(methods of difference and similarity)，普沃斯基和图纳(Przeworski & Teune, 1982)把这两个方法分别称为"差异最小化系统设计"和"差异最大化系统设计"；维巴(Verba, 1967)的"学科框架案例比较法"(disciplined-configurative case comparisons)，该方法与乔治在1982年提出的"结构化重点比较法"(structured-focused comparisons)类似；当然也有其他用到"其他条件不变"(ceteris paribus)假设和反事实的研究方法，等等。上述方法经常被学者援引，但有效使用它们却存在一定程度的困难。由于缺乏有效的对应方法，在一些情况下，定性研究者无法确定由于忽略某个变量所导致的后果——也就是所谓的"遗漏变量"问题。本节将讨论一些有用的对应策略。

我们首先使用语言逻辑讨论遗漏变量导致的偏差，然后辅之以公式化表述。本节的最后将讨论一些存在遗漏变量问题的研究设计。

5.2.1 评估遗漏变量导致的偏差

研究者要估计解释变量 X_1 对被解释变量 Y 的作用。在实证分析过程中，假设 X_1 对 Y 的作用是 β_1。对 β_1 的估计可以选择回归方程或者其他方法，假设其估计值是 b_1。如果采用定性方法，研究者同样可以对该因果关系进行估计，只不过估计过程使用语言逻辑，同时也依赖于研究者的个人经验。

假设研究者试图估计竞选国会议员的支出对候选人得票率的影响。在完成了定量或者定性估计工作之后，一位同行指出估计过程

遗漏了一个重要的变量 X_2，继而使得估计出的作用其实是虚假相关。这位同行指出，如果遗漏的变量 X_2 是“候选人是不是现任国会议员”的话，β_1 的估计值 b_1 是不准确的。她建议模型应当控制这个变量的作用。

那么，该如何评价她的建议呢？需要我们弄清楚的是在什么条件下，忽略该变量会影响对竞选支出作用的估计；而又是在什么条件下，忽略该变量不会影响我们的结论？如果“竞选人是不是现任议员”对被解释变量没有影响，那么遗漏该变量并不会影响对竞选指出作用的估计；也就是说，如果 X_2 不影响 Y，那么它就不会导致估计值出现偏差。这是遗漏变量不影响估计结果的第一个特例：不相关的遗漏变量不会引起估计偏差。因此，如果竞选人是否现任并不影响选举结果，研究者就完全可以忽略该变量。

第二种遗漏变量不会导致偏差的情况是遗漏变量与模型已有的解释变量之间不相关。就上面的例子来说，如果“竞选人是否现任议员”与现有解释变量“竞选支出”无关的话，遗漏前者就不会导致估计结果偏差。因此，当遗漏的变量与研究者感兴趣的主要解释变量无关时，控不控制遗漏变量都不会影响对主要解释变量的估计。之所以要控制某个变量，是因为我们有理由相信该变量和现有变量之间存在相同的差异。研究者需要考虑的不是遗漏变量是否影响被解释变量，而是遗漏变量是否和模型已有解释变量相关。如果不相关，则不必担心遗漏这些变量会导致估计结果产生偏差。①

一旦某个遗漏变量不符合上面两个情况，也就是说，该变量和已有

① 这里读者应当注意这两个特例间的区别。在第一种情况中，遗漏变量与被解释变量无关，估测结果是无偏的，此时我们对被解释变量的预测也不会受影响。但在后一种情形下，遗漏变量虽与被解释变量不相关但与解释变量相关，此时对因果关系的估计仍然是无偏的，但对被解释变量值的预测会受到影响。所以，如果是否现任与竞选支出无关，遗漏是否现任这个因素将不会影响对竞选支出与得票率作用的估计。但如果研究的目标在于预测，我们就希望尽可能地捕捉到被解释变量中所有的系统性变异，那么遗漏是否现任这个重要的因素就会影响预测的准确性。即使研究的长期目标旨在对选举进行全面系统的解释，但在单个研究中弄清楚决定因素的因果效用也不是一件容易的事情。所以，不管长期目标是什么，研究者都应该将精力集中在一个（或少数几个）因素的因果效应上。

解释变量相关的话，那么遗漏该变量就会导致估计结果产生偏差。拿上面的例子来说，同行的担心显然是正确的，因为遗漏的那个变量和解释变量及被解释变量都相关：现任议员赢得的概率一般比较大，同时其竞选花销也较多。

以上论述的公式化表述如下，读者可以参考公式(5.5)的最后一行：

$$E(b_1)=\beta_1+F\beta_2 \tag{5.3}$$

该式用来表示估计 X_1 对 Y 作用时所产生的偏差，其中，F 表示 X_1 和遗漏变量 X_2 之间的相关程度。①如果解释变量只有 X_1，由此估计出该变量对 Y 的作用 b_1 就是无偏的：平均起来，b_1 等于 β_1。在上述两个情况中，误差项 $F\beta_2$ 等于 0，因此估计是无偏的。公式(5.3)就以公式化的形式表述了获得无偏估计结果的条件。如果用公式化表述上面两个容许遗漏变量问题存在的情况，则如下：

- 遗漏变量不影响被解释变量（即 $\beta_2=0$，此时并不需要考虑该遗漏变量和被解释变量之间是否相关）；
- 遗漏变量与已有被解释自变量不相关（即 $F=0$，此时不需要考虑 β_2 的取值）。

在实际研究过程中，一旦怀疑有遗漏变量的问题，研究者就不能轻易忽视。此时应当尽可能控制该变量。如果数据不可得，至少也应当估计出导致偏差的方向。估计出方向非常重要，因为只有研究者了解是高估还是低估了真实因果效用，才能判断已有结论是被强化还是弱化了。

让我们再举一个例子，基于撒哈拉以南非洲国家的研究发现，专制政权下更容易发生政变。

该研究问题的解释变量是专制程度，被解释变量则是政变发生的概率，而样本是位于撒哈拉沙漠以南的非洲国家。当然研究者也可以将样本拓展至整个非洲国家以检验上述假设是否依然成立。如果用 β_1 来表示因果作用，专制政权导致政变概率提高就意味着 β_1 为正数。

① 更精确地说，F 是 X_1 对 X_1 作回归所估计到的系数。

上述因果关系中没有考虑经济条件对政变的影响。虽然没有关于经济方面的数据，我们仍然有理由作出这样的假设：失业有可能增加政变概率($\beta_2>0$)，同时失业率与专制政体似乎存在正向关系($F>0$)。如果进一步假设经济发展水平对政变的作用先于专制程度的话，上述没有考虑经济因素的估计结果便会出现估计的严重偏差。在这种情况下，失业率和解释及被解释变量都是正相关的(意味着 $F\beta_2>0$)。遗漏该变量的估计结果将不再是纯粹专制程度的作用(β_1)，而是两个变量的整体作用($\beta_1+F\beta_2$，而非 β_1)。此时，上述三个变量间的两两关系意味着整体作用大于单一的专制程度作用($\beta_1+F\beta_2>\beta_1$)，即 $b_1>\beta_1$。因此，如果没有考虑失业率，将导致对专制程度作用的高估。在这里，读者需要区别该问题和测量误差导致后果之间的区别，遗漏变量导致的偏差有时候也会把一个负向作用错误地估计成正向作用。

遗漏变量问题并不总是导致因果效应被高估。比如在一些国家，专制程度与失业率是负相关的(F 为负)：这些国家的政府通过镇压，有效地控制敌对派系并维持和平局面，同时也解决了大部分民众的就业问题。此时，依然假设失业率和政变概率是正相关，那么 $F\beta_2$ 就是负数。忽略经济因素会导致专制程度对政变概率作用的估计值($E(b_1)$)低于真实作用(β_1)，这意味着低估了专制程度的作用。在这种情况下，如果 F 为负且 β_2 足够大的话，实际为正数的 β_1 就完全有可能被估计成负数。研究者因此得出完全错误的政策性含义：专制可以减少政变发生的概率！上述论述告诉大家，即使研究者没有失业率的数据，了解这些知识也能让我们避免得到错误的结论。

上述例子说明，在很多情况下，研究者并不是非要通过回归分析来估计参数、评估偏差方向及获得结论的。定性研究和直接观察获得的结论都可能受到偏差的影响，这一点和定量分析相同。本节的内容告诉读者，即使依赖于数据以外的信息，研究者仍可以了解偏差的方向及程度。

研究者有时知道自己的研究设计中存在遗漏变量问题，但却没有这些变量的信息。在这种情况下获得的结论是有问题的，同时很可能会被你的同行发现。那么如何处理呢？假设研究者仍然能从其他渠道

获得这些变量的相关信息。比如，虽然有些变量的信息对于全部样本来说不能获得，但对于部分样本而言确实可得；或者该领域已有的研究提供了关于该变量的部分信息，等等。这些都能为研究者所用。总之，即使这些信息不完全，也有助于我们判断因果效应估计中偏差的方向及程度。

当然，即使研究者已经了解遗漏变量问题带来的偏差，识别出遗漏变量依然不是一件容易的事情。对于这个问题，没有很有效的解决方式，但我们还是建议所有的定量和定性研究者在遇到这个问题的时候，尽量寻找到遗漏变量并把它们纳入分析中。以下的内容将给读者提供一些基本的指导。

即使研究者掌握了那些变量的信息，遗漏变量问题还是会给估计工作造成困难。有时候，学者基于掌握的信息认为这些遗漏变量和被解释变量之间存在正向关系，便利用这种两两变量间关系去分别估算它们的因果效用。这种方法在小样本研究中特别受一些学者青睐，因为同时考虑所有变量会使得估计结果不精确甚至使得研究设计变得不确定。关于这点，我们已经在本书的 4.1 节讨论过。但问题是，在考虑两两关系的时候依然会存在遗漏变量问题，所以这种方法并不能解决所有问题。我们认为，理想的解决办法仍然是研究者尽力搜集到这些变量的信息并在分析中加以控制。就定性研究者来说，小样本研究并不能解决遗漏变量问题，识别偏差的来源更为重要。总之，在研究中如果没有考虑其他影响因素，研究结论依然会存在偏差。

当然读者也不能把我们的建议理解成：如果担心遗漏变量导致偏差，就把所有那些与解释变量相关同时影响被解释变量的因素都纳入模型中。我们的原则是，如果一个变量是关键解释变量的作用结果，那么研究者在分析中就不应该控制这样的变量。为了理解这个表述，让我们看一下这个例子：研究者试图研究 1 万美元的额外收入（解释变量）对选民投票给民主党候选人（被解释变量）的作用。那么需不需要控制一个变量叫“在投票前五分钟询问该选民是否准备投给民主党候选人”？如果控制该变量，那么被解释变量自然会受到影响，同时该变量也可能和解释变量相关。我们的回答是“不应该控制”。如果真的对

选民投票前的决策进行控制，那么原模型所估计的收入对民主党得票率的作用几乎全部来自该控制变量了。因此该变量的作用并不是一个额外需要控制的因果解释。所以如果盲目地应用上述指导原则，就会导致研究者错误地选择控制变量，虽然此时该变量确实影响"是否投票给民主党候选人"，同时也和解释变量收入相关。我们的意思是研究者如果滥用上述原则，其试图考察的关键解释变量的作用就可能被这个变量吸收掉。

为了便于读者进一步了解，让我们再举另外一个例子。假设研究者试图检验原油价格的急剧增长对公众关于能源短缺问题看法的作用。该假设的关键解释变量是原油价格，该信息可以从报纸上获得。被解释变量(公众对能源短缺的感知程度)可以通过民意调查获得。有些读者自然会产生这样的疑问：是否应该控制那些和能源问题相关的电视报道呢？电视媒体关于能源短缺问题的报道与解释变量(原油价格)相关，同时也会影响公众对该问题的看法。但要知道，电视对该问题的报道可能恰恰是原油价格变化产生的结果之一。所以研究者在评估石油价格对公众态度关系时不应该控制该变量。如果研究者关注的是电视报道的作用，就应当控制石油价格，这是因为价格往往先于电视报道(此时的关键解释变量)。①

因此，如果研究者试图估计某一个解释变量总效用的话，就应该根据理论模型将所有影响被解释变量的因素都考虑进来。我们再次重复一下变量取舍的原则：如果某个解释变量是由关键解释变量所引起，那么该变量就不应该被控制。除此之外，对于那些与被解释变量及模型已经考虑变量相关的因素，研究者则应当控制住它们以消除遗漏变量

① 研究者应当了解在控制电视报道之后，油价上涨对公共态度作用背后的意义。以下两个方面的讨论都能加深我们对此因果关系的理解。首先，此时的估计结果反映的是原油价格对公众态度的直接作用，这种直接作用是价格总体作用剔除价格通过电视报道影响公共态度的间接作用后获得的。也可以将这种作用描述成在电视报道不变的情况下，能源价格对公共态度的影响。对于后者，假设研究者可以控制电视台关于能源短缺的报道。这意味着即使油价出现浮动，电视台也不会因此增加或减少报道。此时电视报道就变成一个常量了。虽然在实际操作中，研究者无法实现这样的控制，但是用多元回归的方法通过控制电视报道这一变量，依然可以获得原油价格对公共态度的条件效用。

带来的偏差。[1]

那些由关键解释变量所导致的变量不需要在模型中控制，这一观点对突出理论在研究设计过程中的作用非常重要。理解了这一点，读者就会明白使用模型的目的是建构出理论而非进行“数据挖掘”。“数据挖掘”的意思是研究者随意选取解释变量进行回归和定性分析。如果没有理论模型指导，研究者便无法确定研究应当考虑哪些潜在解释变量。在实际情况中，完全可以出现一个非常显著的结果来自一个无关紧要的解释变量。例如，选民在投票前五分钟的选举意向和其他一些控制变量当然会显著影响其最终选举行为。没有理论作为基础，研究者很有可能不清楚是否需要控制那些彼此相关的控制变量。这个时候，研究者就会犯遗漏变量的错误，同时也容易纠结于那些很琐碎的问题。

在很多情况下，适当地增加解释变量并不是一件很容易的事情。相较那些模型已经控制的变量，需要增加的变量数量并没有一个标准，而资源往往是有限的。一般来说，解释变量越多，意味着模型中任意单个变量的解释力就越小。增加变量的一个目的是避免遗漏变量偏差：如果一个相关变量没有被放进模型，那么该模型的解释力相应地就会受到限制。

关于遗漏变量偏差的公式化分析

让我们先考虑只有两个解释变量的模型：

$$E(Y) = X_1\beta_1 + X_2\beta_2 \tag{5.4}$$

假设在没有控制 X_2 的情况下估计了 X_1 的效应。这种做法在什么情况下没有问题？在什么情况下会造成估计偏差？为了回答这些问题，我们在这里用公式来表述存在遗漏变量时模型的估计系数。当 X_2

---

① 另外，研究者感兴趣的也许正是关键解释变量的直接或间接效应，甚至是方程中包含的其他变量。在这种情况下，合理的做法是基于相同数据进行多次回归。当然，这种做法的前提是研究者能够解释这些回归间的差异。

被遗漏时，变量 β_1 的估计值如下：

$$b_1 = \frac{\sum_{i=1}^{n} X_{1i} Y_i}{\sum_{i=1}^{n} X_{1i}^2}$$

为了确定该估计结果是否存在偏差，让我们根据公式(5.4)对估计系数 b_1 取期望值，可得：

$$\begin{aligned} E(b_1) &= E\left(\frac{\sum_{i=1}^{n} X_{1i} Y_i}{\sum_{i=1}^{n} X_{1i}^2}\right) \\ &= \frac{\sum_{i=1}^{n} X_{1i} E(Y_i)}{\sum_{i=1}^{n} X_{1i}^2} \\ &= \frac{\sum_{i=1}^{n} X_{1i}(X_{1i}\beta_1 + X_{2i}\beta_2)}{\sum_{i=1}^{n} X_{1i}^2} \\ &= \beta_1 + F\beta_2 \end{aligned} \tag{5.5}$$

其中，$F = \frac{\sum_{i=1}^{n} X_{1i} X_{2i}}{\sum_{i=1}^{n} X_{1i}^2}$ 表示 X_1 对 X_2 做回归获得的斜率系数。上述公式最后一行和公式(5.3)相同，此处不再详细讨论。

5.2.2 遗漏变量偏差的例子

本部分将讨论关于遗漏变量问题的几个例子。这些例子既有定量研究，也有定性研究。其中有些例子来自真实研究，有些则仅仅是假设的。我们都知道，教育水平是决定政治参与程度最重要的指标之一：那些受过高等教育的民众更有可能参加选举或其他政治活动。假设研究者从数据中发现这两个变量间存在紧密的相关关系，便想进一步了解它们之间是否存在因果关系。如果有的话，探究教育背景是如何影响政治参与程度的。

研究者需要首先弄清楚是否存在一些变量，这些变量先于后期获

得的教育程度,但同时也会影响教育和政治参与程度。现有研究发现,受访者的种族及父母的政治参与程度等因素属于这样的变量。一般来说,如果父母积极参与政治活动,孩子自然会受到影响。同时,父母也显著影响孩子们的教育。如果不考虑父母的作用,教育程度和政治参与程度之间的关系就很可能是伪相关,或者被严重高估。

同样,受访者的种族背景也发挥了相似的作用:黑人可能会被同时剥夺受教育和政治参与的权利。在这种情况下,数据反映出来的关系并不是真实的因果关系。遇到诸如此类的问题,最理想的情况就是研究者将所有同时影响教育和政治参与程度的变量都考虑进来。

如果控制了这些先行变量(antecedent variable),教育和政治参与间的关系就很可能减弱甚至消失。比如,如果控制了种族,教育可能就不再影响政治参与了。即使是这样,也不能说明教育程度没有任何作用。如果对黑人的歧视会分别影响教育和政治参与程度的话,在这种情况下教育和政治参与之间就没有因果关系。如果种族是通过影响受访者教育程度来影响政治参与的,同时种族歧视将降低黑人受教育的机会,那么教育依然是影响政治参与程度的主要因素。此时,教育对政治参与作用的大小会随着考虑了种族因素而减弱,但并不意味着教育这一变量重要性的减弱。此外,通过以上的分析,读者也可以了解种族和教育是如何相互作用进而影响政治参与程度的。

但是从上面提到的两种情形获得的结论是不同的。如果黑人的政治参与程度较低是由于其教育程度较低所导致,那么我们就可以预测当黑人平均受教育水平提高时,其政治参与率也会相应地提高。但如果黑人不那么积极参与政治活动是由歧视造成的,提高他们的教育水平并不能改善其积极性。

有时候,研究者希望寻找到那些与教育存在共线性或随其变化而变化的变量。前者即那些能够使得教育—政治参与变成伪相关的遗漏变量,比如智力,一般来说智力高的人教育程度也往往较高,同时在政治活动中也较为活跃。后者是那些用来揭示教育程度如何影响政治参与的变量,比如受访者是否修过公民教育或基本沟通技能的相关课程,是否参加过学生自治组织,等等。当模型控制了这些变量之后(此时教

育对政治参与的独立作用就不存在了),如果发现教育程度的作用减弱,这并不意味着教育对政治参与没有作用。比如在模型中控制了受访者是否修过沟通技能课程,而该技能的学习是通过学校获得的同时在不同教育水平上呈现差异,这样的分析过程就会告诉我们教育是如何影响政治参与的。

上述例子再次说明了模型的重要性:在开始数据分析之前,研究者必须先确定好理论模型并在此基础上选择变量。这是指导研究者合理选择变量唯一的正确方法。就上一个例子来说,在实证分析之前,研究者就应当基于理论解释好教育是如何影响公民活动的,进而确定选择哪些变量。尽管定性研究并不需要在回归方程中引入额外影响结果的变量,但就决定是否考虑其他因素这一点看,其和定量研究没有实质性差异。在本书之前的章节,我们提出过这样的研究问题:超级大国之间的峰会如何影响它们的合作?假设研究者发现美国和苏联之间在峰会结束后的合作明显比之前频繁,此时如何就此认定该关系是因果关系而非由于遗漏变量导致的伪相关呢?进一步来说,如果证明是因果关系,如何解释作用机制呢?

读者可以再次思考一下该例子中的先行变量:那些与峰会召开相关同时也直接影响两国之间的合作行为。如果两国领袖对彼此都有信心,他们就会经常会晤从而促进两国的合作。当然还有另外一种可能性:双方在地缘政治方面的野心都受制于国内政治,两国领袖也会安排会晤并寻求合作。如果是后一种情况,虽然峰会昭示两国外交关系朝好的方向发展,但峰会本身并没有促进合作的作用。此外,也可以把峰会理解为因果链上的一环,就像种族影响教育程度,而教育程度影响政治参与一样。当两国领导人相互信任时,峰会就成为彼此加强这种信任的策略,进一步的合作才能由此展开。在这种情况下,就不能说峰会对合作没有作用,实际情况是,峰会和领导人彼此信任之间的互动创造了合作。如果模型同时考虑了信任和野心因素之后仍然发现峰会发挥了独立作用,研究者便可以得出结论:峰会确实促进了合作。下一步的问题就是思考该关系背后的机制。研究者可以通过比较不同峰会,识别出哪些因素有效地促成了会后的合作,哪些没有促成。可能成为影

响因素的变量有：会前准备是否充分、会议议题是经济方面还是国家安全方面、会谈时双方国内情况是否稳定、峰会时的天气情况以及会议用餐情况，等等。对于这些因素的考察需要理论指导，弄清楚它们和解释变量之间的关系并根据逻辑及经验提出假设。

样本数量会显著影响遗漏变量偏差。一般来讲，如果样本数量很少的话，遗漏变量偏差很难避免。一方面，模型加了太多不相关的控制变量，研究设计会变得不确定（见本书 4.1 节的讨论），估计也会变得低效。与之相反，如果模型遗漏了相关的控制变量，就会造成估计偏差。而让研究者棘手的是在研究开展之前无法确定是否控制某个变量。

在这种情况下，研究者很容易下结论说利用小样本进行因果推论是不可能的，但我们恰恰认为获得的结论越有限，对结论就越乐观。如果研究者认识到利用小样本进行有效因果推论是困难的，研究者就会对获得的结论十分谨慎。正如本书第 2 章提到的，好的表述及描述推论也比一个错误的因果推论好。对于定性研究者来说，如果他们把精力放在作出有效的描述推论上，而不是放在那些证据不充分的因果推论上的话，研究质量将会有很大的提升。如果理论能被清晰地建构且可以联系能被观察的结果，研究者就依然能在理解因果关系方面取得进展。上面的描述略显复杂，为了便于读者理解，让我们再看一个关于国际关系研究的例子。

海伦·米尔纳（Helen Milner）的研究《抵制保护主义》（*Resisting Protectionism*，1988）旨在回答一个让人费解的问题：20 世纪 20 年代与 70 年代的美国存在很多相似之处，但为什么只有在后一时期，美国才奉行保守主义的贸易政策？作者提出的假设是：国际依存度在 20 年代到 70 年代之间不断加深，因此美国在两个时期的对外政策是不同的。就研究的样本来说，该研究结论所依赖的观察值只有两个，这无法使作者把她的假设和其他解释区分出来。如果研究到此为止的话，该研究建立的理论就充满了不确定性。因此，作者必须增加案例。

米尔纳的解决办法是详细说明因果效应的作用机制。根据她的假设，那些资本主义民主国家间的经济依赖程度会加强产业及企业对特定政策的游说力度，而游说如果获得成功，就会影响整个国家偏好。基

于该假设,米尔纳研究了美国 20 年代和 70 年代、法国 70 年代的众多产业,她发现那些具有巨额跨国投资及出口导向的产业最支持开放的贸易政策。基于这些发现,米尔纳建立了关于美国政策偏好在 20 年代和 70 年代之间差异的理论。她采用的方法和我们在本书中建议的关键部分是一致的,即合理选择样本并制订研究计划,即使这些样本不是研究者主要的关心对象,基于它们获得的推论也可以用来检验理论。从这个意义上来说,米尔纳的研究堪称典范。

如果说米尔纳的研究有什么不足之处,潜在遗漏变量问题应该是最严重问题。最明显的遗漏变量是企业面临进口产品的竞争程度:竞争越激烈,企业越倾向于保护主义政策。也就是说,面临进口产品的竞争程度和米尔纳试图研究的解释变量有关且在大部分情况下是先行变量。如果该变量再和被解释变量、跨国投资和出口依赖相关的话,米尔纳的研究结果就有可能出现偏差。而根据比较优势原理,进口竞争和出口依赖之间可能存在负相关,因此如果研究没有控制进口产品竞争程度的话,遗漏变量偏差就会出现。

为了解决这个问题,米尔纳只选择那些受进口产品影响严重的行业作为样本进行研究。这样就使一个重要解释变量——进口产品的竞争程度——保持不变,进而可以消除或者至少大大降低了遗漏变量所导致的偏差问题。当然,她也可以只选取那些受到进口产品影响一般的行业作为样本,这样做和上面的逻辑相似:只要保持该变量在所选样本中是常量即可。

尽管控制了进口产品的竞争程度,米尔纳的研究仍然面临其他遗漏变量问题。基于已有理论和实证文献,她认为最值得注意的有两个问题:(1)美国国家力量的变化可能导致其在 20 世纪 20 年代和 70 年代的政策偏好存在差异;(2)美国国内政治的变化同样会影响这种差异。她在研究设计中尝试对第一个因素进行控制:由于美国在国际贸易中的份额在这两个时期大致相同,她便从国家政策、产业政策以及公司政策三个角度来控制美国的国家力量。但米尔纳却忽略了美国在 20 年代奉行政治孤立主义而在 70 年代的国际社会中发挥了领袖作用。上述地位的变化值得进一步研究以排除它们对估计结果的潜在影响。

对于国内政治的变化，米尔纳通过直接比较这两个时期的行业和企业来控制。因为国内政治属于宏观环境，同一时代的所有企业和产业均面临相同的政府结构及政治过程。此外，作者又额外研究了法国20世纪70年代六个面临进口商品竞争的产业。基于法国的研究虽然不能控制住美国国内政治的变化，但基于此，米尔纳发现出口依赖对贸易保护政策偏好的影响并不会随着国内政治的不同而改变。通过上述研究设计，米尔纳充分考虑了几种潜在的遗漏变量问题，进而大大地降低了估计偏差。

但米尔纳的研究还是存在遗漏变量问题。作者的研究对象是"企业的贸易偏好"。我们知道诸如舆论、意识形态、工会、国内政治结构等因素均会影响该偏好，但米尔纳并没有直接控制这些变量(1988：15—16)。从理论上来说，她的做法是站得住脚的，只要这些因素和关键解释变量(出口依赖和跨国投资)不相关或它的部分结果，或这些因素不影响企业和整个行业的贸易保护倾向，是否遗漏它们就不是一个严重的问题。但如果这些因素和它们之间确实存在某种联系，并且在因果顺序上先于(而不是晚于)关键解释变量的话，作者就必须进一步调整研究设计以明确控制这些因素。①

就取样来说，米尔纳采用的方法也存在一定问题，这可能导致其获得的因果推论较为低效。前文提到，她通过控制住进口产品的竞争程度来避免该变量引起的测量偏差问题。从理论上来说，竞争程度在因果顺序上要先于关键解释变量并与其相关，同时该变量也影响研究的被解释变量。米尔纳选择那些面临最激烈竞争的行业作为样本后，并没有再利用其他指标对其进一步区分。她做的是直接对样本中的每个行业和企业进行逐个分析，以期获得国际贸易依存度对其贸易保护偏好的作用。

① 米尔纳对遗漏变量偏差的可能性作了说明，但她的说明并不完全正确："我在研究设计时，刻意将不同时间、不同国家的行业选为观察值，之所以这样做是为了能够确保(遗漏的控制变量)会有变化。这样做之后基本的论点仍然成立。"(1988：15)然而，将"控制变量设为常量"的唯一方法是让它们保持不变，而非使它们有变化。如果理论认为这些变量确实重要，她可以让关键解释变量(国家、行业或企业的出口依赖程度)有变化的同时让那些控制变量不变。

米尔纳的这种样本选择方法对于因果估计来说效率不高，因为关键解释变量在样本中的差异较小(Milner, 1988:39—42)。虽然该取样方法并不会导致严重的推论问题，却意味着米尔纳需要引入更多的案例以期获得更确定的结论(见本书6.2节的讨论)。我们的建议是选择那些使得解释变量差异较大的样本便可以达到这个目的。例如，米尔纳如果在保证相同程度的经济和竞争压力的前提下，再用是否涉及国际业务这一指标进一步选取样本，那么她的估计结论就会更加有效。

研究者永远无法完全排除遗漏变量导致的偏差。在米尔纳的研究中，她的研究设计使得她可以逐步处理那些遗漏变量，从而使结论变得更强、更有说服力。米尔纳的严谨研究说明，基于定性材料的社会科学家完全不必担心研究推论的有限性。完美的研究当然是可望而不可即的，同时也没有明确的标准。但谨慎地联系理论和模型能够使得研究设计既可以提高论点的合理性，也能够降低推论的不确定性。

5.3 包含不相关的变量会导致估计低效

上一节内容论述了遗漏变量所导致的估计偏差问题，读者便轻易地认为研究者应当搜集并估计所有潜在解释变量的作用。在开始本节的讨论之前，我们必须声明这并非5.2节的观点。5.2节的结论是，没有控制一个和解释变量不相关的变量并不会导致估计偏差，不管该变量是否显著影响被解释变量。同时，如果一个变量是现有解释变量导致的结果，控制该变量也是不对的。读者应当正确理解上节的内容:研究者不必将所有影响结果的变量都搜集起来，当然也不能轻率地否定那些存在遗漏变量的研究。

确定哪些先行变量可能影响被解释变量并和解释变量相关并不是一件容易的事情。在这种情况下，一旦研究者发现某个因素和解释变量之间存在相关关系，并在理论上能够预期这个因素会影响被解释变量时，研究者可能毫不犹豫地把它包含进模型之中。很多人

认为该种做法是明智之举，可以有效地避免遗漏变量引起的估计偏差。但要知道，符合上述条件的因素可能会非常多，其实它们中的绝大部分是没有必要控制的。这样做也会增加研究结论的不确定性。众所周知，结论的不确定性在小样本研究中是一个非常严重的问题（见本书 4.1 节的讨论）。看似“周密”的研究设计可能会破坏整个研究。本节的内容就是想和读者一起讨论研究者误将不相关变量放入模型中所付出的代价，同时为那些试图“囊括一切”的方法提供限定条件。对于模型包含不相关的变量，我们的观点是：即使该控制变量不影响被解释变量，但如果该变量和主要解释变量相关程度越强，对后者的估计效率就越低。

为了说明这一点，让我们设想以下两种情况中对关键解释变量的估计：第一，除了关键解释变量之外，模型没有包含其他无关变量；第二，除了关键解释变量之外，模型还有一个不相关的控制变量。本节最后部分将提供这两种情况下估计结果和估计效率的公式化表述，这些表述的结论是：首先，关键解释变量的估计结果在这两种情况下都是无偏的。也就是说，即使模型控制了一个不相关的变量，但对关键变量的估计平均上是正确的。其次，如果不相关的控制变量和关键解释变量不相关，那么对后者作用的估计不仅是无偏的，而且是有效的。在实际研究中，如果加入的变量和核心解释变量确实不相关，那么加与不加该变量不会对模型估计有任何影响。但是，如果加的这个变量和关键解释变量高度相关，就会导致估计效率大大降低。

也就是说，如果模型控制了不相关的变量，研究者可能要付出很大代价，估计结果不是真实结果。假设用一个新数据重新估计某项研究，同时研究中有一个不相关变量和关键解释变量高度相关，研究者就很难获得相同的结果，由此会得到一个完全不同的因果推论。如果研究控制了所有不相关的解释变量并保证没有其他错误，那么从中获得的估计结果只是在平均意义上正确：具体到一个研究本身而言，其得到的估计结果可能和真实结果相去甚远。我们要强调的是，对其不断地进行再分析就会获得相同效用仅仅在平均上才有意义。模型中的不相关变量所导致的估计效率降低和研究者舍弃部分

观察值造成的效果相似。这一点就提醒读者：过多地在模型中增加不相关变量相当于对原有样本提出了更高的要求，自然会导致因果推论所基于的信息量降低。

为了让读者清楚地理解，让我们再看一下非洲国家政变的例子。初步研究表明政治迫害程度将增加政变发生的频度。假设有学者对此结果提出质疑并认为该研究没有考虑到这个国家脱离殖民的方式是通过武力还是谈判这一因素。但是我们不同意这位学者的质疑并认为他提到的这一因素并不影响被解释变量的话，模型一旦控制了该变量会导致什么后果呢？

包不包含该因素是否会影响估计结果取决于该因素是否与关键解释变量相关。这个变量测量的是国家摆脱殖民统治的方式，而关键解释变量测量的是政治迫害程度。如果有理由认为这两个变量之间相关，那么包含前者就会严重降低对后者估计的效率。为了便于读者理解这一点，让我们集中到该国实现独立的方式这一因素上。一个可供选择的研究方案是，研究者可以根据通过武力还是通过谈判摆脱殖民统治这一指标把非洲国家分为专制政权和非专制政权。然后，通过分别统计两种政权内部的政变次数来估计迫害程度的作用。在这种情况下，取得独立方式这一因素就被控制住了。虽然这种研究设计可以避免遗漏变量问题，但代价高昂：如果额外的控制变量不影响被解释变量而只和解释变量相关的话，那么采用两个分类就意味着所依靠样本数量的降低，进而导致了估计的低效。毫无必要地降低估计效率无异于抛弃辛苦收集的样本。这样一来，研究者之前付出的很多工作就白白浪费了。针对这一问题，最好的解决方法依然是搜集更多的样本，如果条件不允许这样做，研究者应当谨慎识别出那些不相关的变量并避免在模型中控制它们。

对模型中含有无关变量而导致无效估计的公式分析

假设正确的模型设置是 $E(Y)=X_1\beta$ 且 $V(Y)=\sigma^2$。但是研究者错误地认为还有一个解释变量 X_2 也影响被解释变量，进而将方程设

置如下：

$$E(Y)=X_1\beta_1+X_2\beta_2 \tag{5.6}$$

我们并不知道β_2是否为0，在这种情况下，如果对公式(5.6)进行估计会给β_1带来什么影响呢？

如果Y只受到X_1的影响，其作用的估计值是b_1；如果认为Y同时受到X_1和X_2的影响，此时X_1的估计值是$\hat{\beta}_1$。如果仅从估计结果的无偏性来看(即不断重复估计，每次获得的估计值在平均水平上是否等于真实值)，包不包含X_2都不会影响X_1估计值的这个性质。从这个意义上来说，b_1和$\hat{\beta}_1$都是真实作用的无偏估计：

$$E(b_1)=E(\hat{\beta}_1)=\beta_1 \tag{5.7}$$

但在估计结果的有效性方面，两者就存在显著差异了。正确估计值b_1的方差——见公式(3.9)的计算公式——是：

$$V(b_1)=\frac{\sigma^2}{\sum_{i=1}^{n}X_{1i}^2} \tag{5.8}$$

而$\hat{\beta}_1$的方差是：

$$V(\hat{\beta}_1)=\frac{\sigma^2}{(1-r_{12}^2)\sum_{i=1}^{n}X_{1i}^2} \tag{5.9}$$

$$=\frac{V(b_1)}{(1-r_{12}^2)}$$

其中r_{12}表示X_1和X_2之间的相关系数(见 Goldberger，1991:245)。

公式(5.9)的最后一行表示了两个估计值方差间的关系。如果新加的解释变量和原有的关键解释变量无关，那么方程是否包含前者对后者的估计影响不大。此时，原有关键估计值的方差在这两个情况下是相同的。但如果它们之间存在相关关系，就会显著影响X_1估计值的有效性。一般来说，这种相关关系越高，X_1估计值的有效性就越低。

5.4 内生性

政治科学研究很少是实验性的。这意味着在大多数情况下，研究者只能观察它们而无法操控解释变量。无法操控解释变量所导致的后果之一就是内生性问题。该问题是指在一些情况下出现反向因果问题：解释变量受到被解释变量影响，而不是我们假设的影响被解释变量。在实验中，因果关系的方向是清楚的，因此这种情况很少碰到。但对于定性研究和定量研究来说，内生性是一个普遍且严重的问题。①

如果研究者无法操控解释变量的取值，确定因果方向就会变得非常困难。在定性或定量的非实验性研究中，解释变量和被解释变量的变化通常是由那些研究者无法控制的因素导致。例如国家入侵、政变、通货膨胀、政策、候选人决定参选、选民的选择，等等。在通常情况下，研究者需要将所有因素都拼凑起来，然后再从中确定是谁影响了谁。

让我们再回到美国国会选举的例子。研究者发现，现任国会议员候选人在选举中的优势在20世纪60年代后期以来有显著增加。对于成因，很多人认为现任议员候选人提高了选区的服务水平，比如免费邮件服务、公费旅行预算、设置专人处理选区居民的特定要求、选区建设经费及其他可以利用现任身份为选区谋取的好处，等等。不难理解，很多选民都会因为这些优惠而纷纷把选票投给具有现任身份的候选人。

通过增加选民服务来提高选票比率的假设似乎合情合理，但该假设能获得证据的支持吗？学者们已经提供了非常多的证据来说明该假设是成立的（关于该问题的文献综述见 Cain, Ferejohn & Fiorina, 1987），但相关的实证证据却不多。对于该假设的实证检验需要研究者

① 定性研究者有时也通过参与式观察来调节解释变量的取值。如果研究者系统性地提问不同问题或者在不同的访谈中改变环境，深度访谈也可被视为一种实验。但即便是深度访谈，也会存在内生性问题。这是因为研究者有时是出于便利的考虑而非随机地选择性受访者作为介入组（询问某些问题）。实验方法本身虽然也有很多问题，但内生性问题却不包括其中。

测量出由现任候选人提供的服务水平及获得的相应选票比重，然后再用回归方法估计服务水平的作用。但让人惊奇的是，现有为数不多的实证研究结果都表明该作用为0，甚至是负。

我们认为内生性问题是造成这一矛盾结果的原因。换句话说，恰恰是那些在上一次选举中落选的候选人（不管是由丑闻还是其他原因导致）会提供额外的选区服务。而那些有信心能够再次当选的现任候选人往往将工作重心放在其他方面，比如华府的政策制定等。这样一来就导致了那些提供最多选区服务的现任候选人获得的选票最少。出现这种情况当然不能说是选区服务降低了得票率，只能说对较高得票率的预期减少了现任候选人提供选区服务的动力。因此，如果忽略了反向因果，推论就会出现严重的偏差。

让我们再看另外一个例子，戴维·莱廷曾指出出版于20世纪初期的社会科学经典著作——马克斯·韦伯（Max Weber）的《新教伦理与资本主义精神》（*The Protestant Ethic and the Spirit of Capitalism*）中也存在内在性问题。在书中，韦伯试图证明作为特定经济行为的资本主义精神是新教教义和教旨的意外产物。但作者和他的追随者都无法回答如下质疑：也许正是那些突破前资本主义藩篱的欧洲人才希望摆脱天主教会的束缚。也就是说，特定群体的经济利益也许是导致新教伦理发展的原因。如果没有很好的对照研究，韦伯建立的因果链条可能会被逆转（Laitin，1986：187；R.H.托尼[R.H. Tawney]在1935年就提出了这样的质疑）。

在本节余下的部分，我们将逐个讨论应对内生性问题的五个对策：

1. 纠正推论偏差（5.4.1小节）；

2. 分析被解释变量，并只研究那些是解释变量结果而非原因的观察值（5.4.2小节）；

3. 把内生性问题转化成遗漏变量偏差，并控制遗漏变量（5.4.3小节）；

4. 仔细挑选部分观察值，确保这些观察值中不存在内生性问题（5.4.4小节）；

5. 分析解释变量，并只研究那些完全外生的变量（5.4.5小节）。

上述五种方法既可以用来处理内生性问题，也能弄清楚因果假设。如果因果假设在提出时忽略了内生性问题，该问题最终会成为一个理论难题，研究者必须重新论述解释变量是有可能影响被解释变量的。对于上述前两种方法，我们依然用选区服务的例子来说明。剩下的三个方法会用其他定性研究作为例子。

5.4.1 纠正推论偏差

本节最后一个公式(5.13)的最后一行展示了由内生性问题所导致的估计偏差的方向及程度。为了便于读者理解，我们先把这个表达式放在这里：

$$E(b)=\beta+\text{偏差}$$

如果不存在内生性问题，即偏差项为0，那么b的估计值β在平均意义上是无偏的（$E(b)=\beta$）。一旦存在内生性问题，研究者便无法获得预期的因果推论。在这种情况下，估计结果等于真实结果加上偏差项。内生性问题之所以严重是因为我们一般无法确定偏差的方向及大小：它可能大也可能小，可能负也可能正。这取决于具体的研究问题。值得庆幸的是，即使在研究的第一步中没有避免该问题，研究者依然可以通过确定偏差方向和大小的方式尽量克服该问题带来的偏差。

公式(5.13)说明偏差的大小取决于解释变量和方程残差项的相关程度。这里的残差项是指被解释变量中无法由已有解释变量解释的部分。例如，如果假设之前选区服务会增加选票是成立的，那么对该作用的估计系数(β)应该是正数。再比如，如果预期投票结果对选区服务有影响的话，那么偏差项就为负。换句话说，即使研究者考虑了选区服务对投票结果的影响，选区服务也有可能和残差项相关：对选举结果不看好的现任候选人往往会通过提供更多的选区服务来拉票。在这种情况下，偏差项就为负。对选区服务作用的估计将出现偏差(等于真实结果加上偏差项)。因此，即使该假设是成立的，但由于内生性问题的存在也将导致对真实结果的低估。估计结果可能为0，甚至可能为负值，这

将取决于偏差项的大小。我们可以得到这样的结论：如果没有内生性问题，选区服务对选举结果作用的估计值将低于真实作用。此时，估计结果可以作为真实结果的下限，这意味着该假设更站得住脚了。

综上所述，即使无法避免内生性问题，我们也可以通过估计偏差程度来改善推论。至少这种修正能够帮助确定偏差的方向，进而有机会确定真实作用的上限和下限。当然，在最好的情况下，该方法也可以帮助研究者获得完全无偏的推论。

5.4.2　分析被解释变量

重新定义被解释变量可以有效避免内生性问题导致的估计偏差。内生性问题的存在意味着被解释变量中存在两个部分：被解释部分和解释部分，其中解释部分通过反向因果机制干扰推论的获取。该干扰是通过影响关键解释变量发挥作用的。被解释变量中剩下的部分则是真正需要解释的，它是方程解释变量的函数而不是原因。本小节介绍的方法正是通过将被解释变量中真正的被解释部分识别出来以避免内生性问题。

让我们仍然考察选区服务的那个研究。对于该研究的内生性问题，金(1991a)的方法是将样本中将那些完全由现任议员身份而获得的选票部分区分出来。现有文献发现，该作用在最近几年大约能多获得8%至10%的选票数，而现任议员候选人一般可以获得总选票比重的52%。运用统计方法，金首先估计出现任议员身份对选票的作用，该作用可被视为被解释变量中真正的被解释部分。然后作者在该作用的基础上估计出选区服务对选票的作用。由于在选举中，在任身份带来的优势只占全部选票的很小部分，不足以影响现任候选人通过选区服务来增加得票率。通过上述方法，金的研究就成功地避免了内生性问题。该文的结论表明，现任州议员在选区服务上的预算每增加1万美元，那么在下次选举中该候选人的得票数可以增加1.54%(上下限分别是1.58和1.50)。金的这个研究首次为选区服务增加投票的假设提供了实证证据。

5.4.3 将内生性问题转化成遗漏变量问题

避免内生性问题的另外一种方法是将其视作遗漏变量问题。下面让我们看一下比较选举研究中的一个著名问题。长期以来，魏玛共和国的衰败及纳粹在20世纪30年代的崛起困扰着早期的政治学研究者。基于魏玛共和国的案例研究，有学者认为导致其没落的主要原因是魏玛宪法中规定的选举方式是比例代表制。在该制度下，那些代表着特定意识形态、利益集团及宗教团体的小党派也可以在议会中谋得席位。与单一选区制或赢者通吃制度不同，候选人在比例代表制下不需要为赢得选举而进行妥协。因此魏玛共和国的议会中充斥着大大小小的政治团体，它们不愿意也很难在一起工作。在这种僵局下，像国家社会党这样的政治党派便有机会攫取权力（对于该假设的经典论述见赫门斯[Hermens] 1941年的研究）。

上述理论在关于该议题的重要研究中都有详尽的论述。历史学及政治学领域的学者也普遍认为小团体可以在议会中获得议席及不愿意彼此妥协是导致魏玛共和国衰败的重要因素。在我们看来，该假设依然存在不少问题。当然，对于如此复杂的问题用单一因素进行解释存在这样或那样的问题本身不足为怪。在本节中，我们只把讨论的重点放在内生性问题上。首先，上述解释背后涉及一系列因果推论：比例代表制使得那些缺乏竞选基础的小党派也可以在国会中谋得议席。读者知道，国家社会党正是由此成功进入国会的。国会无休止的争吵导致国内政治陷入僵局。民众对此不满就为后来国家社会党的政变提供了诱因。

但是针对该议题的后续研究认为，国会中政党的分化局面并不是由比例代表制导致。和上述论点不同，有研究者认为，也许正是由于这种分化的局面才导致采用比例代表制度而非其他选举制度。如果是这样的话，导致魏玛共和国衰落的原因就不是选举制度而是政党分化的程度。利用相同因素去解释其他结果（在本书第1章中，我们提到研究者应当在数据中为假设去寻找证据而不是制造证据），研究者就会发现

当社会中的政治团体对其他民族、种族或者特定宗教群体的态度非常狭隘偏激时，这个社会更倾向于采用比例代表制度。因为这是唯一能被社会各派都接受的选举制度。读者如果回顾一下实行比例代表制之前的德国历史，通过溯清各个党派的起源就能理解该观点。虽然比例代表制度使得众多小党派在国会中占据一席之地，但我们并不能由此就认为是选举制度导致了党派分化的局面。政党分化虽然不是比例代表制的唯一原因，但该选举制度的采用和议会中呈现的分化局面确实是社会分化的结果（对于该论点的早期研究见 Lakeman & Lambert，1955:155）。

该观点意味着内生性问题转化成了遗漏变量偏差问题。也就是说，魏玛共和国早期存在的社会分化现象是遗漏变量，该变量导致了比例代表制的采用。相对于选举制度，社会分化在因果链条中属于前因。该因素同时导致了魏玛共和国后期的衰落。通过这种转换方式，研究者便可以对该遗漏变量进行测量并在进一步的研究中对其控制。当这个变量被控制后，研究者便发现比例代表制度和魏玛共和国衰落之间的关系不存在了。

和早期研究相比，对选举制度和民主间关系的研究已经取得了显著进展，但对该研究问题的争议一直就没有平息过。研究者早期基于的案例研究并不太注重解释的逻辑，后来学者对此问题的研究大都采用多样本分析，并在此基础上逐渐解决变量的测量问题从而完善了因果推论。多样本分析使得研究者可以采用系统方法将内生效用剔除掉以获得解释变量的外生效用。

5.4.4 对样本进行选择以避免内生性问题

在研究政治思想对政策的作用时，内生性是一个普遍存在的问题（Hall，1989；Goldstein & Keohane，1993）。一般来讲，政治家的政治思想往往受到其所处环境的影响，其物质利益很大程度上也是其周围物质环境的反应。那么对政治思想作用的研究工作就会受到遗漏变量物质利益的影响：物质利益同时影响政治家的政治思想和政治策略（见

本书5.4.3小节)。如果将政治思想视为政策的合理化表述,那么前者仅仅是后者的结果而非原因。在这些情况下,政治思想就是内生的了:表面上可以用来解释政治家的策略,其实恰恰是这些策略所导致的后果。

对于检验政治思想对政策作用的研究来说,在方法论上最大的挑战是处理遗漏变量和内生性问题对推论的影响。同时,在实际情况中这两个问题又紧密联系在一起。如果研究者试图检验政治思想的作用,他就必须证明政治家的思想切实影响其施政决策,而不是施政决策及物质利益的简单反映。对于物质利益,研究者需要仔细地加以识别。此外,本问题的被解释变量(政策)和解释变量(政治思想)之间的关系必须结合反事实环境:对持有不同政治思想个体的施政纲领也要加以考察。

对于政治思想内生性或外生性的区分来说,比较研究是一种很有效的方法。比如,尼娜·哈尔彭(Nina Halpern, 1993)的研究试图考察领导人的政治思想在苏联之外的社会主义国家中对于是否采用斯大林模式的作用。作者假设东欧及中国的领导人普遍信奉斯大林模式。该假设解释了这些国家在第二次世界大战之后的经济政策。当然,这些国家领袖对斯大林模式的信奉及实施和哈尔彭的分析确实一致。但该一致性并非就代表这两者之间存在因果关系。正如上文提到的,政治思想可能是内生的:斯大林模式会导致信奉该模式的思想出现并捍卫这种模式,同时施行该模式的预期也会导致这种政治思想的产生。

虽然并没有用本书的语言,但哈尔彭针对此问题采用的处理方式和本书5.4.3小节介绍的方法非常相似:将内生性问题转换为遗漏变量偏差问题。对此,她提出了另外一个替代假设:东欧及亚洲共产主义国家在第二次世界大战后实施计划经济仅仅是受苏联军事及政治的影响。基于这个假设构建反事实是:即使东欧和中国的领导人不信奉斯大林计划经济的理念,这些国家仍然可能施行计划经济并由此产生相应的政治思想。

在研究了那些第二次世界大战后被红军占领的东欧国家之后,哈尔彭认为这些国家之所以采用计划经济模式并不是基于对斯大林模式

的信奉，而是慑于苏联的强权——“这些国家选择计划经济完全是遵守斯大林的命令，该替代假设很难被推翻”(1993:89)。作者于是选取那些不存在该遗漏变量的样本：中国和南斯拉夫。这两个国家都是在第二次世界大战后没有被苏联占领的社会主义国家。前者幅员辽阔同时正在进行内战，斯大林无法施加命令；后者的共产党人在没有苏联红军协助的情况下取得政权，其领袖马歇尔·铁托并不听命于莫斯科。

即使没有苏联政治及军事的统治，中国和南斯拉夫在第二次世界大战之后也都采用了斯大林模式。基于这些事实，哈尔彭认为苏联影响并不能完全解释政策选择。此外，也有学者认为中国之所以采用苏联模式，是由于国内经济状况和苏联非常相似。但哈尔彭认为这个假设一样存在内生性问题。

在排除了政治思想受苏联或者国内经济情况的影响之后，哈尔彭便认为中国(以及南斯拉夫在一段很短的时间内)实施斯大林模式可为研究那些刚经历革命存在不确定性的国家提供基础。由于可供研究样本非常有限，哈尔彭的分析依然比较初步。但我们完全有理由认为，在某些特定情况下政治思想可以是外生的并对政策发挥重要作用。

这个例子告诉读者，首先要把内生性问题转变为遗漏变量问题，在此基础上再有目的地选择那些不存在遗漏变量问题的子样本。哈尔彭就通过这个方式成功地做到这一点。这样，她就可以有效地在主要假设和备选假设之间做出比较。哈尔彭用多个备选假设来解释社会主义国家采用计划经济模式的原因，哈尔彭发现只有用中国和南斯拉夫作为研究对象才能合理地将斯大林模式(即研究关注的政治思想)视作外生。因此，作者将研究重点放在这两个国家上。假设没有上述谨慎的研究设计以排除内生性问题的干扰，作者就不能获得如此让人信服的结论。读者可以自己思考一下，如果她选择波兰和保加利亚这两个国家，结论还有说服力吗?

5.4.5 分解解释变量

本小节将介绍第五个，也是最后一个避免由于内生性问题导致偏

差的方法。该方法旨在将可能的内生解释变量分成两个部分：一部分是纯粹外生的，而另一部分则有一定的内生性。在因果分析中，研究者就可以只用外生的部分来避免内生性问题。

维巴、施洛茨曼(Schlozman)和布雷迪(Brady)研究过在政治活动中自由参与的问题。该研究的目的就是回答是否参与的内生性：尽管社会地位都比较低下，但为什么非裔美国人比拉美裔更热衷于参加政治运动？作者发现有很多因素可以解释这种差异，比如移居美国时间的长短及语言能力方面的差别。但这篇研究关键的解释变量是是否参加宗教活动(礼拜或犹太集会等)。研究者如果没有控制宗教参与情况，那么就无法排除潜在的内生性问题。事实上，人们之所以参与宗教活动，正是因为这些人是政治活动中的活跃人士。教堂其实是一个有浓厚政治氛围的场所，人们有机会在那里学习参政技巧，这正是很多热衷政治的人经常去教会的原因。带有政治背景的神职人员可为教会成员提供关于政治活动的基本培训，同时能激励他们参加政治活动。换句话说，在这个因果关系中，真正的因果方向是从政治活动到宗教活动而不是相反。

针对这个问题，三位作者的应对方法就是分解解释变量。他们提出宗教机构可以通过两个渠道影响成员的政治参与程度：第一，人们通过这些机构举办的活动学习参政技能(例如，发表演讲或组织会议)。这些技能的获得使得人们更加乐于参与政治活动，同时也更加胜任在政治生活中的角色。其次，在参与宗教活动的过程中，人们会获得参与政治活动的激励(例如，对某个政治事件的讨论或接受相关机构邀请参加某项政治活动)。这些激励会显著影响人们参加政治活动的概率。对于上述两个渠道，作者认为前者属于外生因素，后者则带有一定的内生性。也就是说，宗教因素通过后者影响个体参与政治活动部分取决于成员本身是否积极(研究的被解释变量)。

下一个问题自然是通过一系列研究来检验上述分类法是否合适。作者首先论证不同类型的教会如何影响成员学习参政技巧。作者认为，天主教会是一个高度阶层化的组织：神职人员由上级教会委派，普通教会成员没有机会参与教会的管理。因此，参加天主教会活动的成

员学习参政技巧的可能性并不高。和天主教教会相比,新教的教会成员较多地参与教会管理,因此其学习参政的机会要多得多。在美国,大部分非裔美国人都信奉新教,而拉美裔则大多信奉天主教。作者由此认为,正是这种不同导致了两个群体在政治参与技巧获得上的差别。虽然基于这两个群体,但该文的发现依然可以拓展到解释其他种族的政治参与。例如,对于美国白人来说,参与宗教的不同同样意味着在参政技巧上的高低。这种差异其实和人种无关,当然也不受被解释变量(政治参与程度)的影响。

在确定参政技巧获得渠道的外生性之后,作者直接将其测量出来并将其设为该研究的解释变量。这样就取代了之前具有内生性的"是否参加宗教活动"这一变量。通过上述方法对解释变量进行分解并将研究重点集中在外生的部分上,作者们成功地解决了内生性问题。

在这个例子里,解决内生性问题的关键是非裔及拉美裔美国人所属的教会并受政治参与行为的影响。事实上,属于哪个教会一般在成员未成年之前就跟随家庭确定下来了。不过从更长远的时间看,非裔美国人对新教的信奉可以追溯至奴隶制时期及南方的种植园制度。而拉美裔美国人信奉天主教是因为西班牙对拉丁美洲的殖民。同样,天主教和新教在教会制度上的差异当然不是由其神职人员参与当前美国政治生活的兴趣所导致。这种差别来自 16 世纪宗教改革运动,对此作者也进行了充分论证。

对内生性的公式化分析

以下我们将用公式演示内生性问题造成的估计偏差。假设我们要估计解释变量 X 对被解释变量 Y 的作用,并用如下公式估计:

$$E(Y)=X\beta \tag{5.10}$$

该公式也可以写成 $Y=X\beta+\varepsilon$,其中 $\varepsilon=Y-E(Y)$,是误差项或者干扰项。进一步假设解释变量存在内生性——Y 在一定程度上会反过来影响 X:

$$E(X)=Y\gamma \tag{5.11}$$

如果公式(5.10)是正确的,但在估计过程中却忽略了公式(5.11)所表现的反向因果关系,那么对β的估计会造成什么样的影响呢?换句话说,在错误地认为$\gamma=0$的情况下,β的估计值是:

$$b=\frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2} \tag{5.12}$$

为了检验该估计值的一致性,我们需要计算它的期望值如下:

$$\begin{aligned} E(b)&=E\left(\frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2}\right) \\ &=E\left(\frac{\sum_{i=1}^{n}X_i(X_i\beta+\varepsilon_i)}{\sum_{i=1}^{n}X_i^2}\right) \\ &=\beta+\frac{\sum_{i=1}^{n}C(X_i, \varepsilon_i)}{\sum_{i=1}^{n}V(X_i)} \\ &=\beta+\text{偏差} \end{aligned} \tag{5.13}$$

其中,偏差$=\sum_{i=1}^{n}C(X_i, \varepsilon_i)/\sum_{i=1}^{n}V(X_i)$。如果$X$和误差项$\varepsilon_i$之间的协方差$C(X_i, \varepsilon_i)$为0,对$\beta$的估计将不存在偏差。此时,$b$的期望值等于$\beta$,$b$是无偏的。在大多数情况下,研究者可以假设模型未能解释的部分ε和解释变量X之间是不相关的。但在这个例子中,在考虑了解释变量X的作用之后,X中依然有一部分差异来自被解释变量Y的反作用。因此,内生性问题意味着公式(5.13)最后一项中的第二部分不为0:β的估计值是不一致的。

因为X方差始终是整数,因此偏差的方向就取决于上文提到的协方差。当然当X的方差非常大并超过协方差时,偏差并不突出,但是这种情况并不常见。读者可以自己回顾一下本节所讨论的例子,特别是关于偏差项的解释。

5.5 对解释变量进行赋值

本书4.4节指出优秀的控制实验需要实现两点:控制样本的选取及控制解释变量的取值。我们当时只讨论了第一点。本章之前的部分也分析过了遗漏变量偏差及其他方法论上的问题。本章以下部分将讨论第二点:如何对解释变量进行赋值。

在医学实验中,药物和安慰剂被随机分配给病人。这种对样本的随机赋值在大样本研究中非常有用,但在样本较小的情况下优势并不明显。在大样本研究中,对解释变量的随机复制可以消除其内生性问题(这种方法使得解释变量不会受到被解释变量的影响)及测量偏差问题(只要我们能对介入进行准确测量)。当然,对大样本的随机赋值,最重要的是使得遗漏变量问题得以克服,因为关键解释变量的赋值是随机的,因此不大可能和遗漏变量相关,即使是和那些可以潜在影响被解释变量的遗漏变量也不会相关。但是对于小样本来说情况就不同了:对解释变量进行随机赋值依然不能保证其和某些遗漏变量相关,而这种相关性会导致估计结果出现偏差。样本选择偏差的例子就告诉我们,随机赋值的变量依然可能和被解释变量相关。与此类似,如果样本数量较少的话,即使是随机赋值的解释变量,也非常有可能和遗漏变量相关。

在实验研究中,学者可以对解释变量赋值,但对采用定性方法的研究者来说就没有这么好的条件了。如果研究者自行选择解释变量的取值或者有其他因素影响其取值的话,就会造成样本选择偏差或者内生性问题,同时估计效率也会大大降低。例如,假设研究者试图通过实验方法检验游行对参与者政治效能感(political efficacy)的影响。普遍采用的方法是随机安排某些个体参加游行,另外一些人则待在家中。然后比较两组实验群体在效能感上的差异(也可以比较两组人参加实验前后效能感的差异)。但事实上,是否参与游行往往由参与者自行决定。该决定可能会受到个体特质(比如年龄、是不是学生等)及其他因

素的影响(比如学校距离游行线路的远近等)。而所有影响个体是否参与游行的因素很有可能也和被解释变量的政治效能感相关。

让我们再举一个例子,研究军备发展对战争可能性的影响。这个问题是国际关系领域的经典问题,同时该问题的研究单位是国家且样本数量较为有限。一个国家军备预算会增加其后续参与战争的可能性吗?在该问题中,解释变量是军备预算(可以用其预算开支占 GDP 比重或者军备预算数额的变化来测量),被解释变量则是有无战争:在其后特定时间段内是否有战争发生。理想的实验设计当然是对解释变量进行赋值:选择不同的国家作为样本,然后自主决定每个国家的军备预算(通过随机赋值或者用下面将要讨论的"有目的"地赋值方法)。读者都知道这显然是不可行的。在实际情况中,研究者一般采用这些国家政府公布的军备开支作为解释变量的取值。但读者应当明白,官方数据的取值不大可能独立于被解释变量。也就是说,如果一个国家为了应付即将到来的战争而增加军备支出的话,该问题就会受到内生性问题的影响。该内生性问题意味着反向因果:国家对未来战争可能的预期会影响当期的军备支出。在研究联盟和战争的关系时,内生性问题也困扰着研究者。一些国家选择结盟,但研究者却不能操作这些国家结盟然后再检验这种结盟与否对战争有影响。结盟本身并不能被视作外生的解释变量,它同样受到对未来战争预期的影响。

上面这些例子告诉读者,内生性问题并非在任何情况下都存在,不过一旦变量是内生的,研究者就必须面对。一般来说,弄清楚变量的赋值过程非常困难,同时也没有特定的解决办法可以依循。尽管如此,这依然是研究者不得不面对的任务。

随机取样及赋值在实验中造成偏差的概率会随着样本数量的增加而迅速下降。因此,即使研究者手头的样本不是太多,也完全可以采用随机方式进行选择和赋值。当然,如果样本数量"足够多",那么随机取样可以完全满足本书 3.3 节所讨论的有条件独立假设。至于多少观察值才算是"足够多",我们将在 6.2 节中给出标准。在定性研究中,研究者能够获得的观察对象数量或有信息的观察对象非常有限,随机取样和赋值在这种情况下就不太可行了。对于控制实验来说,如果样本数

量不足，研究者同样无法用这两个方法获得推论。

如果遇到上面的问题，定性研究者应当问问自己是否需要增加样本数量。在无法获得全部样本的情况下，最可靠的方法是随机搜集数据然后再对解释变量进行随机赋值。如果这种方法不可行，就不应该随机选择数据。此时，研究者应当凭借对研究对象的知识积累来选择样本。这些知识可以基于现有研究、那些研究者认为比较可靠的猜测及该领域其他研究者的判断。然后尽量采用可以避免偏差和估计低效的方式对解释变量进行赋值。如果上述方法依然不可行，也就是说研究者知道估计存在偏差但无法判断偏差的方向和大小时，研究者能做的是尽可能降低推论的不确定性。读者在了解了本书关于因果推论的讨论之后，应当可以对此做出自己的判断。我们建议所有的研究都应当专门开辟一个章节来详细解释样本选择及赋值的过程。通过这些解释工作，研究者应该使读者知道他们在样本选择及赋值时采用的规则、可能出现的偏差及相应的对策。

5.6 控制研究环境

如果研究者有目的地选取样本时没有考虑相关控制变量及其他威胁因果推论的问题，那么单位同质性假设就无法满足。我们必须确保解释变量的测量误差尽可能地小、和一些重要的遗漏变量无关以及不会受到被解释变量的影响。也就是说，研究者必须有效解决本章以上内容讨论过的测量偏误、遗漏变量及内生性问题。如果采用了所有的方法依然无法消除这些问题，研究者也应当尽可能去识别、评价并尝试克服它们。

在基于小样本的田野研究中实施控制是非常困难的，但却对避免因果推论的偏差非常重要。然而在实际操作中，定性研究者很少考虑控制变量，有些研究甚至没有任何控制变量。博伦、恩特威斯尔和奥尔德森(Bollen, Entwisle & Alderson, 1993)对该问题的研究综述发现，

在整个社会学著作和期刊文章中，有超过 1/4 的研究者根本不用任何控制变量。

让我们看这个例子，假设研究者试图研究坐牢的经历是否影响个体持有极端的政治信仰。一个比较理想的研究设计是采用实验方法：随机选取个体，再随机抽取一半的人监禁一年。然后再测量两组人宗教信仰的极端程度。估计的因果效应应当是这两组人一年之后在宗教信仰极端程度上的平均差异。如果采用大样本分析，那么有条件独立的假设是成立的，在这种情况下可以获得的因果推论也是无偏的。可是上面介绍的两个方法都不太可行。

如果假设实验的参与者很少，由于 4.2 节讨论的问题，即使采用随机取样及赋值也不能满足条件独立。在这种情况下，研究者需要明确的控制。就上面的例子来说，最简单方法莫过于在实验前先测量出参与者政治信仰的激进程度，然后再通过测量两组人实验前后在程度上的变化来获得因果估计。上述方法可以控制住实验参与者在介入前的差异。为了进一步让读者弄清楚该方法是如何获得因果效应的，我们最好回顾一下因果推论中的根本问题。在理想状况下，实验者先选择一个参加者，通过严格控制使得个体所处的外部环境保持不变，在一年之后对该参加者政治信仰的激进程度进行测量。与此同时，再把这个人送进监狱待一年，然后同样测量其所持有的政治信仰激进程度。那么监禁对于政治信仰的因果效应就是上述两个测量值之间的差。①但因果推论的根本问题是，我们在同一时点只能观察到一种情形下的个人信仰：同一个人不可能同时在监狱内和监狱外。

所谓的控制就是试图以最直接的方式克服上述根本问题的一种方法。既然研究者无法在同一时刻观察到同一个人的政治信仰，那么可以替代的方法就是寻找两个(组)个体，这两个(组)个体除了在关键解释变量——是否入狱——上存在差异外，其他方面都尽量相似。读者需要知道，我们不能按照他们信仰的激进程度选择样本。研究者可以

① 为了严格按照第 3 章提到的步骤，研究者必须重复进行同一试验，然后对每次试验获得的因果效应取平均。此外，个体因果效应的方差也可能成为研究兴趣。

首先选取一批最近刚从监狱释放出来的人作为样本，然后给每个人都匹配另一个人，这个人在很多方面都和那个刑满释放人员相似但唯独没有进过监狱。那如何匹配相似的人呢？研究者可以先采访这些刑满释放人员，了解这些人的性格和生活经历，然后进行匹配：那些在人口学特征上相似，或者在同一社区生活过并在同一学校读过书的人。

再根据某一变量进行匹配后，两个人在该变量指标上应当相同。估计监禁经历对政治信仰的作用时，研究者就需要对其他指标进行控制。由于干扰因果推论的因素有很多，在实际研究中控制其实是一项很困难的任务。如果研究者无法控制某一个能够影响被解释变量的因素，同时该因素又和关键解释变量相关（就是说该变量对政治信仰的影响对于犯人和非犯人来说是不同的），那么由此获得的因果估计就是有偏的。

政治科学在很多情况下是基于国别的研究，在这种情况下，想要通过控制达到研究对象的同质性是非常困难的，因为任何两个国家之间的差异实在太多。比如说，对于很多人来说，比利时和荷兰可能是很相似的国家：根据普沃斯基和图纳（1982）的研究，它们都是经济开放的欧洲小民主国家，同时两者都没有受到邻国的威胁。因此，很多人认为将这两个国家放在一起比较是可行的（Katzenstein，1985）。然而，熟悉这两个国家的人都知道，两者在语言类型、宗教、资源、工业化程度及其他政治方面都存在很大差异。任何一项针对两国政治模式的比较研究中都存在不确定风险。

如果研究者把比利时和荷兰放在一起进行总体比较，推论的不确定性就不可避免。但如果研究者把研究稍微具体化一些，比如欧洲的民主小国的形成是不是殖民地宗主国对其政策的影响，这种情况下研究者完全可以将比利时、荷兰和葡萄牙以及其他那些不是殖民统治历史的国家（如奥地利、瑞典、瑞士和挪威）在一起比较，这种研究设计就比较合理。需要提醒读者的是，除了殖民统治历史，这些国家在其他方面仍然有很多差异。对于单位同质性较为重视的研究者可能直接比较比利时、荷兰及葡萄牙在殖民统治前后的政策差异。这种方案可以有效弥补两种国家的比较研究存在的问题。在这个方案中，比利时并不

是一个独立的观察值，而是通过它实现控制：研究者比较该国在20世纪60年代其殖民地独立前后的国家政策。

但是除了殖民统治的历史，仍然有许多因素没有被有效控制，这些因素使得国家之间仍然存在着差异。对单位同质性非常敏感的研究者可能会考虑另一种研究设计方案，那就是研究比利时、荷兰及葡萄牙在殖民统治前后的政策。作为一个备选项，这项方案可以有效弥补第一种方案的欠缺。在这项方案中，比利时并不是一个独立的观察值，而是通过它进行控制：研究者将20世纪60年代殖民地独立前后这些国家的政策进行比较。相似国家前后比较的理由是拥有殖民统治历史的比利时、葡萄牙及荷兰在政治上的组内差异远远小于和那些没有殖民经验国家的组间差异。因此，通过这种基于时间序列上的对比方法就把组间差异控制住了。事实上，不管是跨国比较还是基于同一国家的前后比较都会面临样本同质性问题。对于前者来说，那些被纳入考察的国家在和研究问题相关的很多方面都存在差异，这些差异没有被测量及控制。同样，即使是同一国家，在不同时点也存在很多方面的不同。当然，这两种差别还是有区别的。不管是跨地区比较还是跨时间比较，没有一种方法能达到完美的控制。但是如果研究者有心将两者结合起来，其提供的证据当然要比只用其中一种更加有力。

研究者需要注意的是，在研究过程中，有目的地选取样本可能会导致严重的问题。该问题在为了控制住潜在相关变量而对样本进行匹配时尤其需要注意。这种方法存在的最大问题是仍然有可能控制不住某些重要的变量。让我们通过下面的例子详细说明这个问题。在这个例子中，研究者采用的正是观察值匹配法。为了促进非洲国家的民主化进程，美国政府设立了一个名为“民主援助”的项目。该项目旨在通过向非洲国家提供诸如宣传民主政治的教材等方式来促进其实现民主化。研究者试图评价该政策对受援国民主化程度的真实作用。在实际情况下，任何一个国家在同一时间只能有“接受援助”或“没有接受援助”这两个状态，它们不能同时发生。针对该问题，研究者采用“预期比较法”(prospective-comparative approach)。也就是说，研究者比较的是那些即将获得援助的国家和不会获得援助的国家。同时，研究者在

这两组国家中再根据所有的相关控制变量进行匹配。匹配的结果是这些国家除了在是否接受美国政府援助这一变量上不同外，在其他方面都较为相似。

研究计划看似完美无缺，但在实际情况中由于语言障碍和时间限制，研究者最终只选择了两个国家进行研究（我们接下来要讨论的问题也存在大样本的研究中）。他首先选择一个接受大量美国援助的国家，再选择一个没有接受援助的国家。然后再测量这两个国家自美国援助开始之后两年在民主化程度上的变化并作为被解释变量。由于有很多因素会同时影响该研究中的解释变量及被解释变量，研究者需要根据这些变量来选择合适的样本，以确保消除遗漏变量问题。

研究者认为应当考虑的两个变量是国民教育水平及是否存在反政府武装。显而易见，这两个变量都和解释变量及被解释变量相关，如果不控制会导致估计偏差（对于遗漏变量偏差的讨论见本书 5.2 节）。但问题是，美国政府一般倾向于支援那些教育系统相对完善的国家（也许是因为这样的国家能和华府建立良好的关系，也许由于美国政府重视教育），而教育水平一直以来是民主的重要决定因素。同时，美国政府也倾向于支持那些没有反政府武装存在的国家。显而易见，反政府派别的存在会降低民主化的可能性。针对这两个变量进行样本匹配，研究者希望可以控制它们对因果推论的影响。

即使如此，依然存在遗漏变量的可能性。如果这些变量同时和解释及被解释变量相关（并在因果链上先于关键解释变量），遗漏它们将会造成估计偏差。研究者遇到的困难在于，如果匹配不恰当或者不完全，就完全有可能遗漏一个和解释变量及被解释变量都相关的变量。

为什么会这样呢？读者应当注意，在匹配过程中研究者必须找一个接受很多援助的国家和一个没有接受援助的国家。假设这两个国家在上文提到的两个变量上都相似：国民教育水平和国内局势较为安定，正如下面所示：

A 国：高援助，高教育水平，和平。

B 国：低援助，高教育水平，和平。

读者自然就会问为什么 B 国那么好却没有和 A 国一样获得美国

的援助呢？想必这个国家有某些“特别”之处吧。该“特别”之处如果同时和解释变量、被解释变量相关的话，遗漏它就会导致估计偏差。比如，如果 B 国政府的军事力量较为强大，这就意味着政府可以有效提高国民教育水平及镇压反政府武装。那么忽略军事力量就会导致有偏估计。如果研究者用低教育水平和高国内暴力局势进行匹配，上面说的问题一样存在。在后一种情况下，我们会认为连这样的国家都能获得美国援助，这也太反常了吧。为了解决这个问题，研究者在中等教育水平和中等国内安定程度的样本中进行匹配。不过即使这样，最终获得的样本在上述两个指标上也会存在不太一致的情况。总而言之，通过匹配获得的样本在控制变量上的取值也许会偏离我们的预期，而这种偏离可能恰恰是遗漏了某些关键变量所导致。

为了便于读者进一步理解，让我们回到之前讨论的监狱例子。研究者试图为受访的刑满释放人员匹配控制个体。匹配后的两个人除了是否有坐牢经历之外，必须在社会经济水平、家庭背景及学校成绩等方面都相似。因此，最成功的匹配策略是匹配的控制个体具备尽可能多的可以导致入狱的因素：来自贫民窟、年少辍学、吸毒、家庭破碎，等等。匹配越是成功，监禁经历和政治信仰之间的关系就越显著。但问题也随之而来：如果对照个体具备那么多可能导致他入狱的因素，为什么这个人实际上没有进监狱呢？是否有其他特别之处？也许他是一个虔诚的宗教信徒？而宗教信仰显然同时和解释变量（监禁）及被解释变量（政治信仰）都相关。

让我们再从另外一个角度理解配对过程中存在的这个隐患。本书的 2.6 节讨论了两种随机差异。正如前文提到，匹配方法的潜在问题是可能遗漏变量。然而，对于基于一大串控制变量匹配获得的两个观察值，研究者仍然有理由怀疑其中一个存在某种“特别”之处，而该“特别”之处没有被研究者所发现。也就是说，依然存在遗漏变量问题。在这种情况下，研究者需要担心的是被解释变量中的随机部分是否影响观察值。由于对变量的测量值很可能远离真实值，在匹配过程中研究者有时刻意去用这些不寻常的样本去匹配。但由于该变量取值如此不寻常以至于找到合适的匹配非常困难，因此上述随机性的存在就增加

了遗漏变量问题的可能。

当然,上面讨论的问题并不能让我们放弃匹配方法。事实上,该方法在小样本研究中非常有价值。只是研究者需要注意,和其他针对小样本的研究策略一样,该方法也存在随机性危险。当然,该风险在大样本研究中并不是问题。我们对此的建议是,通过匹配选择案例,然后在同一案例内部再用其他方法进一步选取观察样本。

采用匹配方法的目的是消除遗漏变量带来的偏差。在实际操作中,致力于比较政治研究的学者经常讨论是尽可能地选择相似的样本(Lijphart, 1971),还是选择差异较大的样本(Przeworski & Teune 1970)。我们认为,不管是"最大相似方法"还是"最大差异方法",都没有回答"相似到底和什么相关"这个问题。研究者经常对此迷惑不解,为之展开的讨论也没有确切结论。其实这两种方法都不是最好的。对我们而言,搜集数据的目的是给因果假设提供最大化的解释力。真正的顺序是围绕这一目的再去选择匹配方法。因此,读者应当跳出围绕方法优劣展开的无谓争论,而将重点放在寻找能回答研究问题的数据上。

在匹配过程中,对于那些可能同时影响解释变量及被解释变量的因素,研究者可以选择那些在该因素上取值相同的样本来控制其影响。比如,在比较美国及其他说英语的前大英帝国殖民地的政治发展时,西摩·马丁·利普塞特(Seymour Martin Lipset, 1963:248)就控制了一系列背景变量。作者认为,"美国、加拿大及澳大利亚这样的前英国殖民地国家都拥有相当开放的大陆边疆,直到今天依然是横跨大陆的联邦国家"。同时,将这些国家作为考察样本也控制了其他因素,比如经济发展水平、政治体制及价值观等。

再举一个例子,戴维·莱廷(1986)在宗教信仰对政治作用的研究中就使用了一种特别谨慎的匹配方法。作者将尼日利亚——一个有着悠久伊斯兰教和基督教信仰传统的国家——作为研究对象。莱廷试图比较这两种宗教传统对该国政治的影响。但我们知道,除了宗教差异之外,两种宗教所在地区还存在其他很多方面的差异。如果忽略了这些因素,就会导致遗漏变量偏差。"伊斯兰教在尼日利亚北部影响巨

大，过去几个世纪以来，这些地区都和伊斯兰世界保持着直接联系，其政教一体的政治要早于英国殖民统治。19 世纪初伊斯兰教复兴圣战之后，正统伊斯兰教就控制了尼日利亚的大片国土。（相比之下）基督教直到 19 世纪末才在该国出现……基督传教士建立的教会学校为该国国民提供西方模式的教育，而资本主义企业家则鼓励当地人种植经济作物，将这个国家逐步融入世界经济体系。"(Laitin，1986:187)

正如莱廷在研究中问道："如何控制诸如民族、经济发展水平、面向世界市场的开放程度、皈依宗教的动机，甚至是生态环境等方面上的差异？这些差异广泛地存在于基督教和伊斯兰教地区。"(1986:192—193)对此，他的对策是在尼日利亚的约鲁巴地区选择一个研究地点。在该地区，两种宗教差不多在同一时间被传入到当地的同一民族，同时当地民族皈依宗教的原因也相似。

至此读者应当可以看到，不管是科利针对印度三个邦的研究、利普塞特针对三个前英国殖民地国家的研究，还是莱廷对约鲁巴地区基督教和穆斯林教群体的研究，他们都没有做到完全的匹配。完全的匹配是无法完成的任务。匹配工作首先要求研究者预测并识别出可能的遗漏变量，然后选择那些在遗漏变量上取值不变的样本来控制它们的影响。需要知道的是，我们永远都无法确定是否已经穷尽了所有可能导致估计偏差的因素。因此，研究者必须对匹配选择过程作出专门评价。通过匹配的方法来实现控制，可以增加有效推论的概率。

综上所述，有两种方案可用研究者选择来获得因果推论：第一，随机取样和赋值。这种方法在大样本研究中较为有效。该方法可以自动实现条件独立。对于实现样本同质性，该方法也比有目的地选取样本来得简单。随机取样和赋值可以避免遗漏变量问题及研究者采用某种和被解释变量相关（在控制解释变量之后）的规则去选择样本所导致的潜在问题。此外，这种方法也可以确保选择样本及获得结论的过程中不会出现人为偏差。第二种方法是有目的地取样，该方法比较适合小样本研究。和基于随机方法的大样本研究相比，在小样本研究中采用该方法获取因果推论往往有一定风险，并较多依赖于研究者对经验世界的判断。同时，控制过程也会带来其他潜在的偏差。不过正如上文

的讨论，在小样本研究中，控制是不可避免。通过匹配使得那些控制变量相同的策略在我们看来是值得尝试的。在此基础上，研究者就可以把注意力只集中到关键的那个解释变量上，从而提高对研究问题的解释力。

5.7 结论

通过第5章的讨论，我们希望定性研究者可以从这些内容中受益。当然，读者在实际研究过程中不必照搬这些方法。一般来说，问题总是扎堆涌现而非逐个出来。比如说，研究者遇到的问题是：不太严重的选择偏差问题，被解释变量有一定的随机测量偏差以及一个很重要的控制变量很难测量三个问题。针对这三个问题，你根据本章的建议逐个处理即可。但对一些更复杂的问题，定性研究者就需要非常谨慎地处理。采用公式化模型表述是一个比较有用的策略，但研究者应当使之与具体研究问题契合。当然很多复杂的数理模型在统计学文献中都有涉及，所以研究者并不需要从头建立模型。

无论采用公式化模型与否，定性研究者都应当重视推论中的方法论问题。在我们看来，定性研究者在获得因果推论的过程中遇到的方法论问题和定量研究者遇到的问题，在一定程度上是相关的。

6 增加观察值数量

在本书之前介绍的内容中，我们一再强调增加理论解释力的重要性。实现这个目标最基本的方法就是找到尽可能多的观察对象并将其转化为研究所用的观察值。我们提到，不管是基于一个观察值的单一案例研究，还是基于有限样本的比较研究，在分析的不同层级都包含很多和研究问题相关的潜在观察值。因此，在无法获得足够样本的限制下，通过尽可能多地增加观察值数量，研究者仍然可以将那些不易上手的研究设计变得易于处理。本章作为这本书的总结部分，我们将向读者提供如何在社会科学研究中增加观察值数量的建议。

我们首先从单样本研究开始，在这些研究中研究者手头只有一个观察值，即 $n=1$。也许有人质疑，如果确实只有一个观察值的话，如何避免"因果识别中的根本问题"呢？我们的答案是：无法避免。但实际情况是，即使在这些单样本案例研究中，研究者仍然可以在样本内发现一些观察值并做比较。一般来说，对这些样本内部的观察值进行严谨的比较研究，研究者依然可以据此作出因果推论。

在本章 6.1 节中，基于单一观察值的研究设计似乎对从事案例研究的研究者不利。由于单一案例中事实上包含很多潜在的观察值，因此单一案例研究过程并非如想象的那么悲观。在对基于单一观察值的案例研究进行评价后，我们将在 6.2 节中进一步讨论获得确定性的推论所需的观察值数量。最后，在 6.3 节中，我们将告诉读者，基本上所有的定性研究都可以转化为包含多个观察值的单样本研究。如果研究者已经正确地概念化已观察到的样本，实现这个目标就并不意味着需要额外的成本去搜集更多的数据。

6.1 基于单样本的因果推论研究设计

在所有研究的研究设计中，研究者面临的最棘手的问题是仅从一个样本中获得因果推论，也就我们所说的 $n=1$ 的情况。在以下内容中我们将对此问题进行讨论，并告诉读者完美地解决这个问题是极端困难的，6.1.1 小节首先介绍哈里·埃克斯坦的那篇论述案例研究的经典论文。然后在 6.1.2 小节，我们用类比推论的方法来分析一个具体例子。

6.1.1 "关键案例"研究

埃克斯坦认为如果无法弄清楚某种特定结果发生的条件，研究者便无法证实或者证伪理论(Eckstein, 1975)。在这点上我们和埃克斯坦的意见一致：研究者致力建构的理论应该能够进行精确预测，并通过来自真实世界的数据对其加以检验。

然而，埃克斯坦在此又进一步认为：如果已经建构出可以做出精确预测的理论，研究者便可以采用"关键案例"来辅助理论并解释问题。在这里，他所谓的"关键案例"研究指的是"对任一变量单独测量"(按照我们的话就是只有一个观察值)的研究。在埃克斯坦的文章中，作者试图传达的主要观点是"案例研究的最大价值就在于对所有备选理论进行检验"(1975:80)。他还特别指出，"单凭一个'关键案例'，研究者便可以淘汰掉某个理论"。埃克斯坦还认为，"关键案例"研究存在这样一种可能性：即使是非常精确的理论，也可以被一个观察值驳倒。需要特别说明的是，按照埃克斯坦的逻辑，如果研究者基于某些理由所使用的案例不太符合理论的预期，但这个案例中最不可能发生的观察结果也没有证伪该理论，那么这个理论就通过了测验。相应地，我们也有充足的信心认为它是对的。与此相反，如果某个理论是否合理仅仅基于那个"最有可能发生"的结果的话，就不能认为该理论通过了检验，它只能

作为一个“合理性存疑”而有待于进一步被检验。

毫无疑问，埃克斯坦的观点非常有价值，尤其是他认为，研究者需要明了对某个理论应该通过“最不可能发生的结果”还是“最有可能发生的结果”来检验。推论的有效性在很大程度上取决于证实或者证伪该理论所进行检验的难度。但埃克斯坦通过关键观察值来检验理论的方法与因果推论的根本问题是矛盾的。如果按照他对“案例”的定义，也就是我们所谓的单一观察值，我们只能说埃克斯坦的上述观点是错的了。①

对于“关键案例”研究是否可以实现埃克斯坦为其设定的理论解释目的，我们对此表示质疑，其理由主要基于以下三个方面：(1)用以解释现实的理论很少有仅依赖于一个观察值的。一般来说，研究者需要若干个观察结果，才能检验大于一个解释变量的作用。(2)对变量的测量很困难同时也不可靠。(3)这个世界并不是由一个必然的决定过程所塑造。因此，即便实现了精确无误的测量，现实世界中种种不确定的因素也会产生随机误差。

1. 替代解释。在开展案例研究时，研究者往往先作一个假设：解释变量可以用来解释某种结果。但随着研究的开展我们发现，那个结果可能是其他原因导致的。研究者在这个情形下就需要将原先假设和新发现的替代解释都纳入考察范围。但此时研究者手头只有一个观察值，因此无法对两种假设进行检验（见本书第 4.1 节）。这种情况下，匹配法也无法解决问题。即使实现了对各个变量的匹配，研究者也至少需要两个值来反映解释变量的变化。总之，在只有一个观察值的情况下，研究者无法检验两种假设。

2. 测量偏误。即使建构出的理论有很好的预测能力，研究者仍然需要警惕另外一个问题：测量偏误。和任何测量工作一样，对理论涉及的变量进行测量时误差也许在所难免（见本书第 5.1 节）。如果研究者手头只有一个观察值，测量偏误完全有可能促使我们误认为一个正确

---

① 但正如我们随后指出的，埃克斯坦似乎意识到了其理论中的缺陷，这就是为什么他最后也提倡用多观察值而非单一观察值去检验理论。

假设是错误的，或者认为一个错误假设是正确的。如果想让理论预测精确，研究者就必须对相关变量进行精确的测量。这个问题在多个观察值时并不突出：研究者可以通过整合的方法减少偏误。总之，如果只有一个观察值，研究者就必须时刻警惕测量偏误所导致的完全错误结论。

3. 确定性。基于单一观察值的因果推论之所以有这样或那样的缺陷，主要是因为真实世界是不确定的。假如真实世界具有足够的确定性，那么推翻假设就变得十分简单：只要真实结果和理论预期不符合。这是因为任何理论都有可能忽略掉一些未知变量。这样一来，即使理论中的基本模型设置没问题，就仍然会出现某些非预期的结果。在这种情况下，研究者无法用这个非预期结果证实或证伪原有理论，因为你不清楚这个结果是不是由于其他因素所导致的。再退一步，即使研究者有一个好的研究设置，在这个设置中研究者手头有两个相同的观察结果。如果改变解释变量的值都会导致其对应观察结果变化，我们就认为解释变量和观察结果之间存在因果关系了吗？在这个充满或然的世界中，也许是一些非系统的随机因素碰巧造成上述变化的吧。如果是这样，造成这样的原因到底是随机因素（见本书第 2.6 节）所致还是研究设计中没有考虑到所有解释变量所致就不再重要了。在上述任何一种情形下，研究者对相关关系的预测工作都应该被理解成在多大程度上准确。对于这部分内容的理解，可以参考埃克斯坦的论述。在他眼中，任何研究都会受到随机因素的影响：

> 所有研究都不能完全排除随机因素对结论的影响。即使是基于大样本的比较研究，在表述其结论时也应该尽量使用诸如"较为可能"这样的字眼。基于一个样本的关键案例研究和比较研究的真正区别在于，后者可以用一些常用的统计方法将结论成立的可能性表示为一个数（例如，显著性水平为 0.05）。这在关键案例研究中是无法做到的。

我们同意埃克斯坦的上述表述：将结论成立可能性进行量化是基于大样本研究的通常做法。在这里我们想补充的是不论观察

值多少，在研究中将随机因素纳入考察范围更是十分必要的。①

综上所述，我们认为基于单一观察值的研究无法有效检验理论或者研究假设。当然，即使我们所进行的研究确实只有一个观察值，该样本对因果推论仍然有价值。比如说有其他人的研究也是基于单一样本，也许这两个研究可以整合成一个：新的研究就有两个观察值。这种策略意味着研究者不应该将解释逻辑和具体的研究过程混为一谈。这样，基于两个不同研究中的观察值，研究者便可以作出有效的因果推论。但这种策略的前提是已有两个研究的样本都较为系统且有可比性。当然，上述策略并不是在任何时候都适用。但对于总结历史细节和描述性推论（见本书 2.2 节）来说，基于单一观察值的研究方法依然十分有用。显然，包含多种观察值的案例研究一般不会受制于此类问题。当然，该类案例研究并不在以上讨论的范围内。

6.1.2 类比推理

如果政策制定者或政治分析人员可以采用类比推理（见 Khong，1992）方法来理解政治事件的话，再利用单一观察值从事研究就不大可取了。类比推论的实质就是在感兴趣的解释变量之外尽量保持其他解释变量在不同样本间没有差异。如果实现了上述设置，那么研究者便可以作出相应的因果推论：不同样本如果解释变量相似的话，那么它们相应的观察结果（被解释变量）也会相似。例如，B 国目前经历的危机和 A 国去年相似，如果两国在其他方面均非常相似且不存在随机误差，那么我们就可以预期 B 国目前的危机造成的结果会和 A 国去年经历的相似。通过这个例子，希望读者可以大概了解这种推理的逻辑。

然而，通过类比推理获得推论还是无法和比较分析媲美。就后者来说，观察值越多意味着推论的确定性越高。因此，尽量增加观察值的策略至少没有坏处。拿上文提到的国家危机的例子，如果能再引入五

① 由博伦、恩特威斯尔和奥尔德森共同开展的比较社会学调查发现，在他们调查的著作和论文中偶然性都发挥了一定的作用，即使是那些有意识地采用穆勒差异法的研究也不例外。

个国家的样本,研究者将随机因素中和掉之后也许会发现,A 国的危机是由随机因素导致的。通过这个例子读者应该看出,类比有点像配比方法。对于后者来说,如果不能保证所有其他因素都相似,那么相应的因果推论就会出错。因此,在社会科学研究中,研究者应当尽可能把握推论的不确定程度,这一点至关重要。在一般情况下,我们都不建议只增加一个观察值做类比,不管它和已有样本多么相似。换句话说,基于汇集大量样本信息的比较法总是优于类比法,即使前者采用的样本并不十分相似。原因非常简单:对感兴趣的观察值会出现的结果,类比法只是利用了另外一个样本的信息,但比较法却汇集了其他大量样本的信息。只要比较法额外搜集的大量样本和试图预测的样本间存在共性,我们就能获得准确而有效的推论。研究者在采用类比法之前,都应该思考一下你的研究问题是否可以采用比较法(见本书 2.1.3 小节的讨论)。①

6.2 到底多少观察值才算足够?

说到这里,也许定性研究者要问一个有关数量的问题:到底多少观察值才算足够?该问题在实质上已经牵涉到对现有研究的评价及对新研究思路的重新设计。对这个问题的回答需要结合所设计的研究方案、要验证的因果推论及其他一些不可控因素。

我们用一个简单的定性模型来回答这个问题。和本书第 4 和第 5 章的公式推导一样,在这里仍然使用线性回归模型并重点关注变量 X_1 的作用。模型中的其他变量均被当作控制变量,这可以避免遗漏变量的偏误问题。那么,研究者需要的观察值数量可由如下公式给出:

① 卡内曼、斯洛维克和特韦尔斯基(Kahneman, Slovic & Tversky, 1982)描述了一种推论中的谬误心理,这种心理发生在决策者在不确定情况下,选择那些时间上最接近或者最容易获得的事件进行类比推论,最终造成系统性的决策失误。他们把这种情况称为"可得性启发法"(availability heuristic)(Keane, 1988)。

$$n=\frac{\sigma^2}{(1-R^2)S_{x1}^2V(b_1)} \tag{6.1}$$

其中，n 表示研究需要观察值数量，其值可由 σ^2、R^2、S_{x1}^2 和 $V(b_1)$ 计算得到。这四个部分各有重要含义并影响获得有效因果推论所需的观察值数量。公式(6.1)推导过程所需要的假设都已经在前文介绍过。①以下按照这些部分受研究者影响的可能性从小到大依次介绍这四个部分：残差方差 σ^2、因果推论不确定性 $V(b_1)$、关键解释变量和控制变量间的共线 R_1^2 及关键解释变量的方差 S_{X1}^2。②

1. 残差方差 σ^2。该方差越大，意味着被解释变量中没有被解释部分的方差越大(见本书 2.6 节的内容)，此时获得可靠因果推论所需的观察值数量就越多。这个不难理解，因为干扰因素多的话，研究者很难根据某个指标来判断所需样本的数量。增加观察值的数量可以提高解释力度以获得系统性的因果效应。

在使用直接类比方法时，估计效率越低，需要的样本数量越多。被解释变量的随机误差问题正符合这种情况(见本书 5.1.2.1 小节)。从研究者的角度来看，该测量误差其实就相当于残差方差变大，因为两者在很多情况下很难彼此区分。因此，残差方差越大(或者估计效率越低)，研究者需要的数据就越多。

虽然研究者无法改变现有的残差方差，但该指标依然非常重要，这主要体现在以下两个方面：第一，研究者对研究越了解，该指标(被解释变量中未解释的部分)就越小。在这种情况下，研究者便不需要引入额外数据了。举个例子，假如研究者对美国独立战争中历次战役的起因和结果都了如指掌的话，那么对某个新解释变量作用的检验就不需要很多观察值(战役)。

第二，需要明确的是，对残差方差大小的了解并不会减少需要收集

① 假设 $E(Y)=X_1\beta_1+X\beta$，$V(Y)=\sigma^2$，不存在多重共线性且所有期望值都取决于 X。

② σ^2 是以所有解释变量 $V(Y|X)$ 为条件时被解释变量的方差。$V(b_1)$ 是 X_1 的因果效应估计值标准误的平方；R_1^2 由所有控制变量对 X_1 做的回归计算得到。而 S_{X1}^2 是 X_1 的样本方差。

的观察值数量，但可以帮助研究者明确推论中的不确定性。如果已知 n 的数量，再结合公式(6.1)中的其他四个变量，我们便可计算出推论的不确定性 $V(b_1)$。也就是说，研究者可以通过观察值数量、残差方差、核心解释变量的方差及其和其他控制变量间的共线性来获得因果推论的不确定程度。

2. 因果推论的不确定性 $V(b_1)$。在公式(6.1)中，$V(b_1)$处在分母位置，由此可以得出，研究者容忍的不确定性越大，需收集观察值的数量就越少。在那些亟待研究的新兴领域内，研究者即使掌握有限数量的观察值，也能做出重大贡献。但在相对成熟的领域中，同样的重要贡献需要推论具备相当大的确定性才能说服同行(见本书 1.2.1 小节的讨论)。

3. 核心解释变量和控制变量间的共线性 R_1^2。如果核心解释变量与其他控制变量无关，那么为了避免遗漏变量问题而加入的这些控制变量将不会影响研究所需的观察值数量。但是如果它们之间的相关性很大，研究对数据量的要求也就越高，此时获得同等推论确定性所需要的样本量就越大。

例如，我们试图研究同一企业内女性雇员是否能与男性雇员同工同酬。由于没有官方途径，研究者只能通过非正式的访谈获取信息。该研究的被解释变量是雇员的年薪。最重要的控制变量是雇员的种族。假设在一个比较极端的环境中，如果所有的男性雇员都是黑人，所有的女性雇员都是白人的话，研究者便无法获得因果推论：当种族变量被控制之后，就不可能区分出性别对收入的影响了。性别在这种情况下就成为一个常量了。这是一个多重共线性和研究设计不明确的典型例子(见本书 4.1 节的内容)。当然这个例子比较极端，在大部分情况下，研究者需要担心的是那种共线性较高但还没到完全的程度。例如，研究对象是 15 个雇员，其中男性中只有一个是白人，而女性全是白人。此时，当种族被控制之后，性别的影响则完全建立在不完全共线性的那个观察值上。

上面的例子就说明，一般情况下，解释变量和控制变量之间的共线性程度越高，浪费掉的观察值就越多。为了达到要求的推论确定性程

度，研究者就需要更多的观察值。该点对于读者的研究设计有重要的指导意义。研究者可以通过选取观察值，将关键解释变量和控制变量间的共线性水平设定在较低水平。因此，对于上文的例子，研究者只要访问足够数量的黑人女性和白人男性，就能降低这种相关性。

4. 因果解释变量值的方差 S_{X1}^2。核心解释变量的方差 S_{X1}^2 越大，所需观察值的数量就越少。

同第三点一样，该点也具有很强的实践指导意义：妥善选择观察值有助于降低对样本的需求量。研究者只需要把注意力集中于选择关键解释变量变动范围较大的样本上。比如说，若要研究教育对社区犯罪水平的影响，研究者选取的样本最好同时包括教育水平很高和很低（核心解释变量）的社区。这意味着获得同样推论确定性所需要的观察值更少，研究者花费的功夫也最少。

上述模型适用的前提是因果关系呈线性。也就是说，解释变量取值越大，相应被解释变量的取值就随之变大（或变小）。如果两者之间的关系为非线性，但大致上仍是单调的（即非递减的），那么模型仍然成立且结果不变。如果因果关系呈现出明显的非线性，那么那些处于中间水平的解释变量的作用就可能完全不同。如果基于解释变量的极端值没有发现因果效应，比如社区教育水平与犯罪之间没有发现相关性，而事实可能是只有中等程度的教育水平才能起到降低犯罪率的作用。当然，并不是所有的研究都会遇到如此情况，但读者在设计研究方案确定假设时需要谨慎。

通过对残差方差、推论不确定性、共线性及核心解释变量方差的关注，研究者可以凭借有限的样本量获得相当的解释力。然而，就本小节的标题“到底多少观察值才算足够?”这个问题而言，依然没有答案。正如模型中展示的那样，观察值的数量取决于四个相互独立的部分，而这四个部分则进一步取决于具体的研究。此外，尽管这个模型为读者提供了一些基本参考，但大多数的定性研究并不完全符合该模型。

越多越好，那到底需要多少呢？在最简单的情况下，即残差方差较小、核心解释变量方差较大、核心解释变量与控制变量不相关且对推论确定性要求不高的情况下，大约 5 到 20 个观察值就足够了。当然，更准

确的答案还要取决于模型的设置及各部分的精确取值。但定性研究本质上就不具备这样的精确性,因此我们可能永远无法找到精确的答案。

不管怎样,研究者可以通过增加观察值的数量来避免这些问题。增加样本意味着搜集更多的观察值。但对于定性研究来说,为了获得更有力的解释,有时候研究者需要重新设计研究思路以获取更多观察值。对于这一点,我们将在接下来的内容展开讨论。

6.3 增加观察值的数量

前面我们提到在有限样本基础上开展研究的困难,也讨论了处理这类问题的一些对策。尽管如此,读者也许已经注意到这些建议大多是“次优”的:仅适用于观察值不多的情况,在应用价值上仍然不及增加观察值数量这一策略。①尽管如此,我们依然认为这些“次优”方案也是很有用的。要知道并不是在任何情况下都能搜集到想要的样本,很多研究问题能凭借的样本非常有限,或者搜集成本过于高昂。在本节中,我们将讨论增加观察值数量的几种方法。这些方法通过重新定义理论的方式增加观察值的数量。在样本有限或者增加样本不可行的情况下,这些方法也许就非常有用。与此同时,其研究对象依然集中在那些能证实或证伪理论的证据之上。此外,数据收集工作完成之后,我们给出的这些策略依然有用。

在本书2.4节中我们讨论过案例的定义。按照哈里·埃克斯坦(1975)的说法,案例是指“可对其任何相关变量进行描述和解释的某个现象”。由于“案例”在社会科学中被使用得太过广泛,我们更倾向使用“观察值”这一术语。观察值是指某个样本的被解释变量的测量值(或该样本所有解释变量的测量值)。观察值是定量研究的基础:研究者收

① 增加观察值数量能够带来的益处已经在很多介绍比较研究方法的文献中有过叙述,李帕特(Lijphart, 1971)的著作就是其中的典型代表。

集它们，作为评估理论的证据。在第 2 章我们提到，对于诸如法国、法国大革命、美国 1992 年大选、伊拉克军事入侵科威特这样的事件，研究者并不需要关注它们的方方面面，而是根据理论需要从中抽象出若干方面并设置成解释变量和被解释变量。然后确定研究单位，并在此基础上获得这些变量的观察值。①

检验理论所用到的信息就是相关变量涉及的一系列观察值。本节讨论的内容是如何增加观察值数量，这需要从研究者试图检验的理论和假设入手。在研究过程中，我们应当这样问自己：所要检验的理论和假设可能有怎样的结果？哪些证据可以检验这些结果？做到这些之后，研究者可以从以下三个方面增加观察值：引进新的样本（6.3.1 小节）、在原有样本基础上采用新的测量方法（6.3.2 小节）及两者的结合（引入新样本的同时对老样本采用新的测量方法，6.3.3 小节）。其中，第一个方法可被视作对假设的重新估计：在原有解释变量和被解释变量的基础上结合新的数据；第二个方法可被视作部分地重新估计：采用新的测量方法获得新的被解释变量，但保持解释变量不变；最后一个方法则意味着在原有理论中拓展新的假设出来，同样采用新的被解释变量。②通过上述方法，研究者便可以从一个单一的"案例"中获得很多相互独立的样本来检验理论，这样的做原因是核心解释变量的测量方式有很多种。③

---

① 我们同意威廉·鲍莫尔（William Baumol，1990：1715）对经济史的观察："很多经济史学家试图对特定历史作出解释时，往往给自己设下了圈套。一位作者致力于描述 19 世纪末促使英国社会转型的'五大因素'，或导致 1847 年欧洲经济萧条的'五大因素'，其实是给自己设置了一个不可能完成的任务。自然科学，即使是已取得众多成就并有了大量累积，依然需依赖于控制实验，以获得某个变量的作用。因此，科学家应有所侧重，而不应该期望对某个复杂现象作出面面俱到的解释。"

② 研究者也可以不改变被解释变量但改变解释变量。在大多数情况下，使用这种策略的目的是避免测量误差。而避免测量误差较为常见的方法是对同一个解释变量进行多次测量。

③ 有时候研究者只是重复此前的研究但没有用新的样本。他们这样做是想知道这些结果可否被复制。当然，不同领域使用的具体方法略有不同。定量研究者会用此前研究过的数据重新进行分析；历史学家则会检验另一个史学家使用过的资源；民族志学者会重听访问录音以确定原始结论是否可靠。这些都是有效的方法，因为科学证据应该是可复制的。但这并不是我们所要讨论的话题，因为这些方法都不涉及新观察值的使用。

6.3.1 变量不变，增加新的样本

通过相同测量方法来获取更多观察值是增加观察值数量的标准方法。该方法意味着研究者基于的理论假说及变量不变，但结合了新的样本。在具体实施过程中，研究者可以通过“跨越空间”和“穿越时间”这两个方法来实现观察值的增加。

通过“跨越空间”获取观察值是指寻找其他类似观察值。例如，在印度之外，再引入巴基斯坦、孟加拉国及斯里兰卡的数据。只要研究者不受时间、资金和技术的制约，该方法还是值得尝试的。科利对印度的研究为此提供了范例（见 5.6 节）。在研究中，由于作者知道被选择的三个邦在解释变量及被解释变量上的局限，为了克服该问题，科利将目光投向另外两个国家：一个是阿兰德（Allende）时代的智利，另外一个是罗伯特·穆加贝（Robert Mugabe）统治下的津巴布韦。前者实施的扶贫政策以失败告终，而后者在科利撰写该著作时刚好上台。穆加贝和他领导的政党所持的意识形态有消除贫困倾向，这点和西孟加拉邦相似。在科利看来，扶贫计划的成功需要三个必要条件，而智利缺少的是组织有序的革命政党。①新增两个样本进一步印证了科利的观点。尽管对于新增的两个案例，作者处理得相对粗略，但这并不影响它们作为检验理论的额外证据。

在操作过程中，研究者也没有必要非要超出考察样本的界限。比如说，如果某个理论基于的样本单位是一个国家，研究者完全可以利用这个国家内部不同的行政层级——邦、县、市、地区等——对其展开检验工作。此时，只要把解释变量和被解释变量同时设置在相应的层级即可。举个例子，假设研究社会动荡理论，该理论将社会动荡和农产品价格联系在一起。研究者此次把考察对象集中在印度上，那么通过考察该国不同地区的方式，就可为分析提供非常多的观察值来检验该关系。在这种情况下，研究者完全不必再寻找其他国家就能获得足够的

① 外部压力也导致了阿兰德的失败，但科利将主要原因归咎于内部问题。

观察值数量。

社会政策领域的学者就经常通过研究同一国家内部的地方政府来探索各种政策的起源。科利对印度三个邦的比较分析是该领域的常见研究方式:在州(邦)、城市及地区间进行比较。我们之前已经提到作者取样的方式其实无法检验政府结构对贫困政策作用这个假设。然而,在引入其他国家信息之后,科利把研究单位下沉到地方政府,通过这个策略,他成功地克服了之前的样本选择问题。在研究中,科利将重点集中到基层长老会——地方政府在区、街区及乡村设立的议会。在政治领袖对扶贫的态度及组织方式上,各地区的长老会存在显著差异。这使得作者可以利用这些地区差异来检验扶贫政策的效果。

用于增加观察值的下级单位并非一定是地理意义上的。对于检验民族国家理论,研究并不一定基于几个国家的数据。研究者可以通过考察同一国家内部的政府结构及决策构架来检验该理论。维巴等学者的最新研究提供了通过非地理单位来增加观察值的例子。在本书 5.4 节介绍的例子中,我们解释了非裔美国人比拉丁裔美国人在教会中学到更多参政技巧的原因:两者所参加的教会性质有差异。非裔美国人大多参加新教教会,而拉丁裔美国人则参加天主教会。如果关于教会性质的假设是正确的话,该理论也可以解释其他类型教会参加者行为的差异。作者发现,和天主教类似组织结构的(基督教)圣公会中的教徒在通过教会学习参政技巧方面与天主教徒非常相似。相比较其他教派信徒,比如(基督教)浸礼会,该教信徒的教育程度及经济条件都较高。因此,这样的差异就为作者的假设提供了进一步的证据。

在用新样本复制原有假设时,研究者必须非常谨慎。也就是说,需要判断这些新证据是否适用于检验假设。适用与否取决于待检验的理论及假设,还取决于样本的属性。如果被解释变量是社会福利政策的话,那么州(省)可以出台该政策就可以被作为研究单位。但如果研究者感兴趣的是关税政策,我们都知道该政策是由中央政府制定的,那么州(省)作为研究单位就不再合适了。同样,如果想了解国家在什么条件下会发展核武器,将观察值设置在印度和巴基斯坦的地方政府上就没有任何意义了。再比如,用印度各邦的数据检验农产品价格变化对

社会稳定的影响是可行的,但要用印度政府机构的数据来检验就行不通了:这种关系不可能发生在政府内部。上述事例告诉读者,是否使用下级政府数据取决于理论本身。这也就是为什么我们建议读者首先列举理论可能导致的现象入手,而不是一开始就搜集大量可能和理论无关的数据。只有明确了理论之后,才能妥善地选取样本。

另外一个增加观察值的方法是"穿越时间"。比如,研究者可以考虑用当今的印度和十年前的印度来检验某个感兴趣的理论。很多案例研究都会涉及对相同假设在不同时间段上的多次检验。

通过在下级单位及时间维度上增加样本对于定性研究非常有用。研究者不必到另外一个国家,不必对一个全新决策展开分析,就能实现观察值的增加,这就解决了小样本研究中样本量有限的问题,但在操作过程中研究者需要注意,新增样本是可以检验理论或假设的。只有这样,新增样本才能给检验过程提供额外的解释力。

另外一个需要读者注意的问题是,尽管新增样本会增加额外的解释力,但并不意味着它们能独立地使用。正如乔治(1982:20—23)所认为的,如果观察值之间彼此不相关,那么新加入的样本对解决问题带来的效力就会打折扣。也就是说,只要观察值之间不存在完全相关,就可以完全用现有数据预测新数据。当然这种情况发生的可能性不大,新数据总会提供一些新的信息。一定程度的相关性不会让理论检验工作失效。因此,以相关信息为基础的新数据虽然提供的信息没有那些完全独立的信息大,但它们对于研究还是有用的。

这一结论至少有两方面的指导意义:第一,在处理那些部分独立的样本时,研究者应当谨慎对待结论的确定性。特别是不能预期它们提供的信息和那些独立样本提供的一样大。第二,对于观察值之间的相关性,研究者也应当仔细分析。相关性的存在往往意味着有一个或一系列变量被遗漏掉。举个例子,研究者对美国某个郡的公民参与度感兴趣。由于存在跨界通勤、人口流动及居民在社会经济政治方面观点相近等因素,相邻的郡之间就有一定程度的相关性。从相邻郡中获得的新样本所提供的信息虽然不及那些相隔遥远、彼此独立的郡提供的多,但它们还是为研究者带来了一些新的东西。

再举个例子,让我们回到之前那个农产品价格波动和社会稳定的例子上。研究者基于印度很多邦的数据对两者的关系展开检验。在每个邦中,我们依次测量农产品价格和社会稳定性。但需要读者注意的是,这些邦并不是彼此独立存在的:被解释变量的取值除了受到本邦解释变量的影响外,也许还会受到其他邦遗漏变量的影响。此外,一个邦的社会动荡可能会像理论预测的那样受到该邦农产品价格波动的影响,可能还受到邻近邦社会稳定程度的影响(只能部分地对理论展开独立检验)。在这种情况下,研究者可以控制对邻近邦社会动荡的报道数量。同样,在时间维度上,前后期的影响也会导致类似问题。

上述例子都是为了说明基于新样本对理论检验工作的复制并不意味着要展开全新的研究。如果新增样本采用的形式和现有样本相同,这些新增样本才有用处。通过本节介绍的方法,那些从事案例研究的学者会发现,找到的新观察值数量可能远超他/她的预期。①

6.3.2 样本不变,使用新的变量

研究者也可以在观察值单位不变的前提下,通过改变被解释变量的方式获得检验理论或假设的额外样本。这种方法之所以有用,是因为它试图寻找相同原因的多种不同结果。研究者一般是从理论或假设着手并提出这样的问题:假设理论是正确的,那么除了要考察的被解释变量之外,解释变量还会引起什么结果呢?这就意味着可能还有其他被解释变量。本书第1章提到,有一种关于恐龙灭绝的理论意味着岩石中需要存在某种特定的化学成分。这就说明,即使一个特定的史前事件,也会产生多过一种的后果。

让我们再回到那个农产品价格波动和社会动荡的例子上。在该研究问题中,我们可以用骚乱的次数作为衡量社会不稳定的指标。即使

① 定量研究者发展了很多有价值的统计技术,用于分析那些在时间及空间上呈现自相关属性的数据。这些技术不但能够解决这些相关问题,也能从数据中进一步提取信息。关于这些研究见 Granger & Newbold(1977)、Anselin(1988)、Beck(1991)及 King(1989; 1991c)。

如此，研究者也应当问自己，如果理论是对的，那么除了这个指标之外，农产品价格的波动还会导致哪些影响呢？这种方式也许会促使我们注意到其他推论：那些受到价格波动影响的其他现象。比如说，某些反常行为——选举行为、商业投资或者移民等。价格波动影响社会动荡的机制也许和影响这些指标的机制相同。

罗伯特·帕特南(1993)在研究意大利地区的社会资本对地方政府治理时就采用了类似方法。在他的研究中，政府治理可以用多个指标来衡量。帕特南采用了12个指标来测量政府表现，这些指标涵盖了决策过程、政策推广及执行等各个方面。此外，作者还采用民意调查搜集各地区民众对政府表现的评价。所有这些指标都被用作理论的预测结果。

如果研究者感兴趣的研究问题是关税政策的话，情况会稍有不同。因为关税政策是由中央政府制定的，因此地方政府就不适合作为研究单位。即使解释的变量在地方上存在差异，比如工农业产品，研究者也应当明确该研究问题的解释变量关税在地方政府层面上不存在差异。不过，如果此时将解释变量换成地方政府代表在关税和贸易法案上的投票行为的话，与此相关的理论就可以用地方政府层面的样本检验了。相应地，研究者也就可以为检验工作找到更多的样本。

6.3.3 增加样本的同时也使用新的变量

在研究过程中，在保持已有的样本及变量的情况下，研究者也可以同时寻找新的样本并使用新的变量。如果试图检验一个全新的假设，即使该假设由之前的理论衍生而来，研究者也需要全新的变量。新理论也许并不适用于之前的样本，它需要的样本可能层级更高或者更低。为了便于理解，让我们仍然用农产品的那个例子来说明。农产品价格波动和社会动荡间的关系也许可以用来解释其他研究对象(如公司或者政府机构)内部的不确定，甚至是个体的行为。因此，在研究过程中，我们应当这样问自己：“如果理论是对的话，那么能否用该理论来解释公司、农业合作社或者个人行为呢？它对确定政府机构的分配政策有

什么影响？对于个体对不确定性的心理学反应及行为有何影响？”

如果研究者没有结果可供观察时，上面提到的这个方法就显得尤其有用。针对核战争的研究就属于这一类——核战争从来没有真正爆发过，因此我们就无法直接观察解释变量对核战争的影响。假设认为敌对国双方都拥有核武器时可以避免核战争的发生。很显然，没有可供利用的观察值来直接检验该假设，但如果把这个假设稍微修订一下，也许我们还是能够找到观察值。比如，在两个国家同时拥有核武库就能防止核战争爆发的前提下，研究者可以观察那些一方有核武器，一方没有的敌对国家并评估核战争爆发的可能性。在这种情况下，研究者就可以找到其他样本并对假设进行直接检验了。

对由最初理论衍生出来新理论的检验，有时需要更低层级的研究单位或完全采用新的研究单位。研究者可以由使用国家层面的样本转向使用地方政府层面的样本，也可以从国家或者省级样本直接转向个体数据——生活在这个国家中的个体行为及决策。不同的理论意味着变量之间的关系不同，而每个变量都会有特定的结果及造成这些结果的机制(Dessler，1991:345)。在设计实证检验方案之前，研究者也需要弄清楚“因果机制”:因果假设涉及的各个变量之间作用关系。但在后续研究过程中，研究者在对除此之外的其他机制的定义及寻找的过程中可能发现用来检验理论的新现象(在本书 3.2.1 小节中，我们讨论了因果机制和因果关系之间的区别)。

上文提到新的社会单位、个体、决策都可能成为新样本。但在决定使用这些信息之后，研究者却发现，新的解释变量无法使用原有的研究单位了。一般来说，作为理论所预测的结果和检验理论所依赖的研究单位应当是同一级别的。比如，检验某个理论的单位是国家，而该理论隐含的假设可以用这个国家中居民的行为及态度检验。如果研究者之后依赖的样本是个体层面的并将把与个体有关的心理、经验和社会地位等因素作为解释变量的话，现有的数据就不再适用于原先的国家样本了。

让我们再回到之前的农产品价格和社会动荡的例子上来。假设农产品价格波动越大，社会动荡就会加剧。研究者可以使用国家或省级

政府层面数据来检验该假设。同时,研究者也可以从该假设中进一步推导出关于生活在该国(省)中某些个体的其他假设。比如,那些受价格波动影响较大的群体,如农作物种植者或购买力较低的群体,参与扰乱社会秩序活动的概率就较大。显而易见,对新假设的检验就涉及对个体的社会疏离感及异常行为的测量工作。

利用文化来解释政治现象的研究也离不开类似个体层面的分析。①韦纳关于印度教育和童工政策的研究就是建立在这样的文化解释之上的。作者认为,印度之所以是世界上唯一一个无法依靠法律强制普及义务教育及禁止童工的国家,其原因就在于该国普通民众及社会精英们共同持有的价值观。但问题是,印度仅仅是一个国家,韦纳的研究会被认为只有一个观察值。作者通过多种方法来规避该问题。首先,韦纳将印度和其他已经普及教育的国家相比较;其次,他又在印度各邦之间进行有限的比较。换句话说,作者采用的方法就是改变研究单位的层级。和上面的例子类似,有关印度文化与政策的假设同样可以推导出关于个人价值观和政治立场的其他假设。而新假设所涉及最重要的个体就是那些制定教育及童工政策的精英。因此,韦纳用来检验他的理论的样本就是个体数据。他对这些精英展开深度访谈以获得他们所持的与教育及童工相关的信仰及价值观信息。至此,关于文化与政治现象的宏观假设就和基于个体层面的假设有效地融为一体了。

上述这种从较低研究层级获得更多信息的方法也可以被用到决策分析中。乔治和麦基翁曾经提出一个叫做“过程追踪”(process tracing)的方法:研究者需要密切观察“将各种初始条件转化成最终结果的决策过程”(George & McKeown, 1985:35)。②根据这个方法,研究者并不是把最终结果(比如国际危机所导致的后果)作为被解释变量,而是重新构建出新的被解释变量。例如,在这过程中的每一个决策、决策者对他人行动和意图的认知等等都可以作为新的变量。这种

① 在社会科学研究中将“文化”作为解释变量颇具争议,但这并非本书的话题。我们对此的一贯主张是即使从文化层面上构建出来的解释,也得符合测量上和逻辑上的要求。

② 康纳德·穆恩(Donald Moon)将该方法称为“理性解释”(rationale explanation)或“推理分析”(Moon, 1975)。

方法经常使用个体层面的数据。这是因为那些把初始条件和最终结果联系在一起的理论常常需要对参与者具体动机及认知的理解。“过程追踪”方法就意味着研究者需要搜集那些符合理论预期且能够揭示决策过程的证据。在操作上，研究者可以采取包括访谈及阅读文本记录的方式来探寻决策者的行为原因。

再举个例子，在国际政治中，国家之间的合作可以通过多种机制实现：积极互惠、军事威慑及共同利益等。每个机制都涉及多个解释变量，它们间的关系也并不相同。这种对某两个国家间达成合作过程的密切研究可以帮助我们识别出最合理的机制。该过程同样会涉及对决策者动机及他们之间沟通过程的观察。

我们认为，“过程追踪”方法可以增加检验理论的观察值数量。①该方法是通过那些可能预测到的现象来增加观察值，并由此将理论和实证分析结合在一起。观察值数量的增加意味着该方法能够克服很多研究中样本有限的困境，使得研究者及读者对社会科学的结论抱有信心。就像上文的例子那样，通过对某个研究单位内部个体态度及行为的分析，可以获得更多的观察值。此外，研究者需要对这些个体特质进行控制，其原因在于它们和事件结果有关系。如果仅关注结果，研究者就会由于观察值数量过少进而遇到遗漏变量偏差及推论不明确等问题。通过对个体态度和行为的考察，研究者便可以对因果机制的有效性进行评估了。

读者需要注意的是，通过这个策略不太可能获得非常强有力的因果推论。因为如果研究者识别出一种以上的因果机制，那么相应的解释变量的解释力就不是特别清楚。不过该方法确实可以检验假设，因为对结果的解释一般会涉及机制的讨论，而这种讨论可以提供证伪假

① 对此，乔治和麦基翁称之为“观察值内的解释”(within-observation explanation)。套用埃克斯坦的话来说，该解释是一种通过重新界定分析单位来增加观察值数量的策略。乔治和麦基翁(1985:36)进一步指出，在案例研究中，“系统的行为不能用一个数据点的信息进行概括，而是用一系列的点或者根据时间绘制出来的曲线来概括”。埃克斯坦(1975)也认为这是一个增加观察值数量的有效方法，因为单一观察值是“对变量中的某个测量结果所进行的解释和报告”。

设所需要的样本。此外，该方法也能让研究者描述出不同机制发生作用的频度，这种描述不光说明了机制发生的条件，也可为后续分析提供基础。

在我们看来，对于建立在较高层级研究单位的假设来说，“过程追踪”及寻找参与者的心理基础是非常有价值的方法。但读者需要注意的是，对于本书用到的最基本的分析逻辑而言，该方法应当被视作延伸而非替代。该方法如果要对因果推论有所贡献的话，还得面对推论中常见的那些问题，比如单位同质性、内生性偏差等。在个体决策者层面，如果想获得有效推论，研究者还必须正确提出及回答研究设计中涉及的这些问题：必须准确测量给定的原因并选择独立于结果的观察值（否则会遇到内生性问题），同时还要保证没有遗漏变量的存在。当然，采用该方法获得的因果机制会使得推论更加严谨：它提供了更多的机会来证伪一个假设。总而言之，“过程追踪”及其他基于较低层级研究单位的方法有助于研究者获得关于因果机制的合理假设。同时，对因果机制的描述也有利于后续的因果推论。即使采用这个方法，研究者依然需要面对因果分析中的那些问题。

6.4 结束语

不管是在理论上还是实践中，定性研究和定量研究都会遇到相同的推论难题。如果那些能够帮助我们理解现实的研究设计可行，就必须遵循科学推论的逻辑。这适用于所有的定量、定性、大样本、小样本、实验型、观察型、历史型、民族志、参与式观察等研究。通过对本章的阅读，读者应当知道，相较大样本研究，在小样本研究中，描述及因果推论的根本问题是很难避免的。本书不光提供了增加样本数量的方法，也讨论了如何从小样本中获得有效的推论。

从事定性研究和定量研究的学者可以通过增加与研究问题相关的信息的方式来提高推论的有效性。实现这一目标可以通过增加观察值

数量(见本书 2.7.2 小节),也可以通过随机取样和赋值的方法。本书的大部分讨论旨在帮助定性研究者提高结论的精确性。与之相关的建议有很多,这需要研究者在实际操作中根据情况选择。此外,有些建议是使用文字表达的,它们并不太适合采用定量研究那种公式化的描述。

如果研究者不愿意使用定量方法研究社会现象,就需要警惕由于研究设计中的方法论失误所导致的推论偏差及低效。很多情况下,这些现象的重要程度要超过定量学者研究的议题。因此,定性研究者的描述及因果性推论理应和定量研究一样合理且可靠。为获得有效的推论,定性研究者需要改变传统上对方法论问题的忽视。在研究设计时需要有更强的自觉意识,并在陈述结论时做到更加具体。当然,读者并没有必要重新推敲那些已经发表了的定性成果以探求其科学依据。如果某个研究者基于两个观察值给出了 12 个因果假设,对于这样的研究,读者和审稿人都没有责任为这种不明确的研究设计辩护。那些能够理解且说明自己分析逻辑的学者总是能够创造出更有价值的研究。定性研究者所面临的方法论问题也是所有科学研究者需要面对的。只有内在逻辑得到充分的理解和遵循后,所有的社会科学研究才能作出有效的推论。

参考文献

Achen, Christopher H. 1986 *Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Achen, Christopher H., and Duncan Snidal. 1989. "Rational Deterrence Theory and Comparative Case Studies." *World Politics* 41, no.2(January):143—169.

Alvarez, Walter, and Frank Asaro. 1990. "An Extraterrestrial Impact." *Scientific American*(October):78—84.

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Boston: Kluwer Academic Publishers.

Barnett, Vic. 1982. *Comparative Statistical Inference*. 2d ed. New York: Wiley.

Baumol, William J. 1990. "St. John versus the Hicksians, or a Theorist Malgré Lui?" *The Journal of Economic Literature* 28, no.4:1708—1715.

Beck, Nathaniel. 1991. "Alternative Dynamic Structures." *Political Analysis* 3:51—87.

Becker, Howard S. 1966. "Whose Side Are We On?" *Social Problems* 14:239—247.

Becker, Howard S., and Charles C.Ragin. 1992. *What Is a Case? Exploring the Foundations of Social Inquiry*. New York: Cambridge University Press.

Blainey, Geoffrey. 1973. *The Causes of War*. New York: Free Press.

Bollen, Kenneth A., Barbara Entwisle, and Arthur S.Alderson. 1993. "Macrocomparative Research Methods." In Judith Blake, ed. *Macrocomparative Research Methods* Palo Alto, Calif. Annual Reviews, Inc.

Cain, Bruce, John Ferejohn, and Morris Fiorina. 1987. *The Personal Vote: Constituency Service and Electoral Independence*. Cambridge: Harvard University Press.

Caplow, Theodore, Howard M.Bahr, Bruce A.Chadwick, and Dwight W.Hoover. 1983a. *All Faithful People: Change and Continuity in Middletown's Religion*. Minneapolis: University of Minnesota Press.

——. 1983b. *Middletown Families: Fifty Years of Change and Continuity*. New York: Bantam Books.

Caro, Robert. 1983. *The Years of Lyndon Johnson*. New York: Vintage Books.

Collier, David. 1991. "The Comparative Method: Two Decades of Change." In Dankwart A. Rustow and Kenneth Paul, eds. *Comparative Political Dynamics: Global Research Perspectives*. New York: Harper Collins.

——. 1993. "The Comparative Method." In Ada W. Finifter, ed. *Political Science: The State of the Discipline*. Washington, D.C.: American Political Science Association.

Cook, Karen Schweers, and Margaret Levi, eds. 1990. *The Limits of Rationality*. Chicago: University of Chicago Press.

Coombs, Clyde H. 1964. *A Theory of Data*. New York: Wiley. Courtillot, Vincent E. 1990. "A Volcanic Eruption." *Scientific American* (October): 78—84.

Dahl, Robert. 1961. *Who Governs? Democracy and Power in an American City*. New Haven: Yale University Press.

Dessler, David. 1991. "Beyond Correlations: Toward a Causal Theory of War." *International Studies Quarterly* 3, no.35 (September): 337—355.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "The Journal of Money, Credit and Banking Project." *American Economic Review* 76, no.4 (September): 587—603.

Diamond, Larry and Marc F. Plattner, eds. 1993, *The Global Resurgence of Democracy*. Baltimore: Johns Hopkins University Press.

Duneier, Mitchell. 1993. *Slim's Table*. Chicago: University of Chicago Press.

Easton, David. 1965. *A Systems Analysis of Political Life*. New York: Wiley.

Eckstein, Harry. 1969. "Authority Relations and Governmental Performance." *Comparative Political Studies* 2: 269—325.

——. 1975. "Case Study and Theory in Political Science." In Fred I. Greenstein and Nelson W. Polsby, eds. *Handbook of Political Science*, vol.1, *Political Science: Scope and Theory*. Reading, Mass.: Addison-Wesley.

Elster, Jon. 1983. *Explaining Technical Change: A Case Study in the Philosophy of Science* New York: Cambridge University Press.

Fearon, James D. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* (January) 43, no.2: 169—195.

Fenno, Richard F. 1978. *Home Style*. Boston: Little, Brown.

Ferejohn, John. 1993. "Structure and Ideology: Change in Parliament in Early Stuart England." In Judith Goldstein and Robert O. Keohane, eds. *Ideas and Foreign Policy: Beliefs, Institutions and Political Change*. Ithaca: Cornell University Press.

Ferguson, Yale H., and Richard W. Mansbach. 1988. *The Elusive Quest: Theory and International Politics*. Columbia: University of South Carolina.

Feynman, Richard P. 1965. *The Character of Physical Law*. Cambridge, Mass.: MIT Press.

Fiorina, Morris, and Charles R.Plott. 1978. "Committee Decisions under Majority Rule." *American Political Science Review* 72, no.2(June):575—598.

Fisher, Ronald A. 1935. *The Design of Experiments*. New York: Hafner Publishing.

Fogel, Robert William. 1989. *Without Consent or Contract: The Rise and Fall of American Slavery*. New York: W.W.Norton.

Friedrich, Carl J. 1958. "Political Philosophy and the Science of Politics." In Roland Young, ed. *Approaches to the Study of Politics*. Chicago: University of Chicago Press.

Fudenberg, Drew, and Jean Tirole. 1989. "Noncooperative game theory for industrial organization: an introduction and overview." In Richard Schmalensee and Robert D.Willig, eds. *Handbook of Industrial Organization*, vol.1 Amsterdam: North Holland.

Garfinkel, H. 1964. "Studies of the Routine Grounds of Everyday Activities." *Social Problems* 11:225—250.

Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." *Political Analysis*, 2:131—152.

Geertz, Clifford. 1973. *An Interpretation of Cultures*. New York: Basic Books.

——. 1983. "Local Knowledge: Fact and Law in Comparative Perspective," In Clifford Geertz, ed. *Local Knowledge: Further Essays in Interpretive Anthropology*. New York: Basic Books.

Gelman, Andrew, and Gary King. 1990. "Estimating Incumbency Advantage without Bias." *American Journal of Political Science* 34, no.4(November): 1142—1164.

——. 1993. "Why Are U.S.Presidential Election Polls So Variable When the Vote Is So Predictable." *British Journal of Political Science*(forthcoming).

George, Alexander L. 1982. "Case Studies and Theory Development." Paper presented at the Second Annual Symposium on Information Processing in Organizations, Carnegie-Mellon University, October 15—16.

George, Alexander L., and Timothy J.McKeown. 1985. "Case Studies and Theories of Organizational Decision Making." *Advances in Information Processing in Organizations* 2:21—58.

George, Alexander L., and Richard Smoke. 1974. *Deterrence in American Foreign Policy*. New York: Columbia University Press.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Kruger. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. New York: Cambridge University Press.

Gilpin, Robert. 1981. *War and Change in World Politics*. New York: Cambridge University Press.

Goldberger, Arthur. 1991. *A Course in Econometrics*. Cambridge, Mass: Har-

vard University Press.

Goldstein, Judith, and Robert O.Keohane, eds. 1993. *Ideas and Foreign Policy: Beliefs, Institutions and Political Change*. Ithaca: Cornell University Press.

Gould, Stephen J. 1989a. *Wonderful Life: The Burgess Shale and the Nature of History*. New York: Norton.

——. 1989b. "The Horn of Triton." *Natural History* (December): 18—27.

Granger, G. W. J., and P. Newbold. 1977. *Forecasting Economic Time Series*. New York: Academic Press.

Gulick, Edward V. 1967. *Europe's Classical Balance of Power*. New York: Norton. Hall, Peter A., ed. 1989. *The Political Power of Economic Ideas: Keynesianism Across Nations*. Princeton: Princeton University Press.

Halpern, Nina, 1993. "Stalinist Political Economy," In Judith Goldstein and Robert O.Keohane, eds. 1993. *Ideas and Foreign Policy: Beliefs, Institutions and Political Change*. Ithaca: Cornell University Press.

Hermens, F.A. 1941. *Democracy or Anarchy: A Study of Proportional Representation*. South Bend, Ind.: University of Notre Dame Press.

Hirschman, Albert O. 1970. "The Search for Paradigms as a Hindrance to Understanding." *World Politics* 22, no.3 (April): 329—343.

Hoffmann, Stanley. 1960. *Contemporary Theory in International Relations*. Englewood Cliffs, N.J.: Prentice-Hall.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945—960.

Horowitz, Donald. 1993. "Comparing Democratic Systems." In Larry Diamond and Marc F.Plattner, eds. *The Global Resurgence of Democracy*. Baltimore: Johns Hopkins University Press.

Huth, Paul. 1988. "Extended Deterrence and the Outbreak of War." *American Political Science Review* 82: no.2 (June): 423—443.

Huth, Paul and Bruce Russett. 1990. "Testing Deterrence Theory: Rigor Makes a Difference." *World Politics* 42, no.4 (July): 466—501.

Hsiao, C. 1983. "Identification." In Zvi Griliches and Michael Intriligator, eds. vol.1, *Handbook of Econometrics*. Amsterdam: North-Holland.

Inkeles, Alex, and Peter Rossi. 1956. "National Comparisons of Occupational Prestige." *American Journal of Sociology* 61: 329—639.

Iyengar, Satis, and Joel B. Greenhouse. 1988. "Selection Models and the File Drawer Problem." *Statistical Science* 3, no.1 (February): 109—135.

Iyengar, Shanto, and Donald Kinder. 1987. *News That Matters*. Chicago: University of Chicago Press.

Jeffreys, Harold. 1961. *Theory of Probability*. Oxford: Clarendon Press.

Jervis, Robert. 1976. *Perception and Misperception in International Politics*.

Princeton: Princeton University Press.

Jervis, Robert, Richard Ned Lebow, and Janice Gross Stein. 1985. *Psychology and Deterence*. Baltimore: Johns Hopkins University Press.

Jones, E.L. 1981. *The European Miracle: Environments, Economies, and Geopolitics in the History of Europe and Asia*. Cambridge: Cambridge University Press.

Johnston, J. 1984. *Econometric Methods*. 3d ed. New York: McGraw Hill.

Kahneman, Daniel, Paul Slovik, and Amos Tversky, eds. 1982. *Judgement under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Katzenstein, Peter J. 1985. *Small States in World Markets: Industrial Policy in Europe*. Ithaca: Cornell University Press.

Keane, Mark T. 1988. *Analogical Problem Solving*. Chichester, West Sussex: Ellis Horwood.

Kennedy, Paul. 1987. *The Rise and Fall of the Great Powers*. New York: Random House.

Keohane, Robert O. 1980. "The Theory of Hegemonic Stability and Changes in International Economic Regimes, 1967—1977." In Ole R. Holsti, Randolph M. Siverson, Alexander L. George, eds. *Change in the International System*. Boulder: Westview Press.

——. 1984. *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton: Princeton University Press.

——. 1988. "International Institutions: Two Approaches." *International Studies Quarterly* 32:379.

——. 1989. *International Institutions and State Power: Essays in International Relations Theory*. Boulder: Westview.

Keohane, Robert O., and Joseph S. Nye, Jr. 1977. *Power and Interdependence: World Politics in Transition*. Boston: Little, Brown.

Khong, Yuen Foong. 1992. *Analogies at War: Korea, Munich, Dien Bien Phu, and the Vietnam Decisions of* 1965. Princeton: Princeton University Press.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. New York: Cambridge University Press.

——. 1993. "The Methodology of Presidency Research." In George Edwards III, John H. Kessel and Bert A. Rockman, eds. *Researching the Presidency: Vital Questions, New Approaches*. Pittsburgh: University of Pittsburgh Press.

——. 1991a. "Constituency Service and Incumbency Advantage." *British Journal of Political Science* 21, no.1(January):119—128.

——. 1991b. "Stochastic Variation: A Comment on Lewis-Beck and Skalaban's 'The R-Square'." *Political Analysis* 2:185—200.

——. 1991c. "On Political Methodology." *Political Analysis* 2:1—30.

Kohli, Atul. 1987. *The State and Poverty in India: The Politics of Reform.* New York: Cambridge University Press.

Kreps, David M. 1990. "Corporate Culture and Economic Theory." In James E. Alt and Kenneth Shepsle, eds. *Perspectives on Positive Political Economy.* New York: Cambridge University Press.

Laitin, David D. 1986. *Hegemony and Culture: Politics and Religious Change among the Yoruba.* Chicago: University of Chicago Press.

Lakatos, Imre. 1970. "Falsification and the Methodology of Scientific Research Programs." In I.Lakatos and A.Musgrave, eds. *Criticism and the Growth of Knowledge.* Cambridge: Cambridge University Press.

Lakeman, Enid, and James D.Lambert. 1955. *Voting in Democracies.* London: Faber and Faber.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference With Non-experimental Data.* New York: Wiley.

——. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73, no.1(March):31—43.

Levy, Jack S. 1985. "Theories of General War." *World Politics* 37, no.3(April): 344—374.

——. "Quantitative Studies of Deterrence Success and Failure." In Paul C.Stern, Robert Axelrod, Robert Jervis, and Roy Radner, eds. *Perspectives on Deterrence.* New York: Oxford University Press.

Lieberson, Stanley. 1985. *Making It Count: The Improvement of Social Research and Theory.* Berkeley: University of California Press.

——. 1992. "Einstein, Renoir, and Greeley: Some Thoughts about Evidence in Sociology." *American Sociological Review* 56(February):1—15.

Lijphart, Arend. 1971. "Comparative Politics and Comparative Method," *American Political Science Review* 65, no.3(September):682—698.

Lindberg, Leon N., and Stuart A.Scheingold. 1971. *Regional Integration: Theory and Research.* Cambridge: Harvard University Press.

Lindberg, Leon N., and Stuart A.Scheingold. 1970. *Europe's Would-Be Polity: Patterns of Change in the European Community.* Englewood Cliffs, N.J.: Prentice-Hall.

Linz, Juan J. 1993. "The Perils of Presidentialism." In Diamond, Larry and Marc F.Plattner, eds. *The Global Resurgence of Democracy.* Baltimore: Johns Hopkins University Press. 108—126.

Lipset, Seymour Martin. 1963. *The First New Nation: The United States in Comparative and Historical Perspective.* New York: Basic Books.

Little, Daniel. 1991. *Varieties of Social Explanation: An Introduction to the*

Philosophy of Social Science. Boulder Colo.: Westview.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press.

Lowenthal, Abraham F. 1972. *The Dominican Intervention*. Cambridge: Harvard University Press.

Mankiw, N. Gregory. 1990. "A Quick Refresher Course in Macroeconomics." *Journal of Economic Literature* (December) 28, no.4:1645—1660.

Martin, Lisa L. 1992. *Coercive Cooperation*. Princeton: Princeton University Press.

Merck & Co., Inc., 1989. *Annual Report*. Rayway, New Jersey: Merck & Co., Inc., 1989.

Merton, Robert K.[1949] 1968. *Social Theory and Social Structure*. Reprint. New York: Free Press.

Mill, John Stuart. 1843. *A System of Logic*. Publisher unknown.

Miller, David. 1988. "Conjectural Knowledge: Popper's Solution of the Problem of Induction." In Paul Levinson, ed. *In Pursuit of Truth*. Atlantic Highlands, N.J.: Humanities Press.

Milner, Helen V. 1988. *Resisting Protectionism: Global Industries and the Politics of International Trade*. Princeton: Princeton University Press.

Moe, Terry M. 1990. "The Politics of Structural Choice: Toward a Theory of Public Bureaucracy." In Oliver Williamson, ed. *Organization Theory: From Chester Barnard to the Present and Beyond*. New York: Oxford University Press.

Moon, Donald J. 1975. "The Logic of Political Inquiry: A Synthesis of Opposed Perspectives." In Fred I.Greenstein and Nelson W.Polsby, eds. *Handbook of Political Science*, vol.1, *Political Science: Scope and Theory*. Reading, Mass.: Addison-Wesley.

Neustadt, Richard E., and Earnest R.May. 1986. *Thinking in Time: The Uses of History for Decision-Makers*. New York: Free Press.

Nye, Joseph S. 1971. *Peace in Parts*. Boston: Little, Brown.

O'Hear, Anthony. 1989. *Introduction to the Philosophy of Science*. Oxford: Clarendon Press.

Ordeshook, Peter C. 1986. *Game Theory and Political Theory: An Introduction*. New York: Cambridge University Press.

Palfrey, Thomas R., ed. 1991. *Laboratory Research in Political Economy*. Ann Arbor: University of Michigan Press.

Pearson, Karl. 1892. *The Grammar of Science*. London: J.M.Dent & Sons, Ltd.

Plott, Charles R., and Michael E.Levine. 1978. "A Model of Agenda Influence on Committee Decisions." *American Economic Review* 68, no.1(March):146—160.

Popper, Karl R. 1968. *The Logic of Scientific Discovery*. New York: Harper and Row.

——. 1982. "The Open Universe: An Argument for Indeterminism." In W. W. Bartley III, ed. *The Postscript to the Logic of Scientific Discovery*. Totowa, N.J.: Rowman and Littlefield.

Porter, Michael E. 1990. *The Competitive Advantage of Nations*. New York: Free Press.

Przeworski, Adam, and Henry Teune. 1982. *The Logic of Comparative Social Inquiry*. Malabar, Florida: Krieger Publishing Company.

Psathas, George. 1968. "Ethnomethods and Phenomenology." *Social Research* 35:500—520.

Putnam, Robert D., with Robert Leonardi and Raffaella Y. Nanetti, 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Ragin, Charles C. 1987. *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Rivers, Douglas, and Morris P. Fiorina. 1989. "Constituency Service, Reputation, and the Incumbency Advantage." In Morris P. Fiorina and David Rohde, eds. *Home Style and Washington Work*. Ann Arbor: University of Michigan Press.

Robinson, William S. 1990. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15:351—357.

Rogowski, Ronald. 1987. "Trade and the Variety of Democratic Institutions." *International Organization* 41, no.2(Spring):203—224.

Rosenau, Pauline. 1990. "Once Again into the Fray: International Relations Confronts the Humanities." *Millenium: Journal of International Studies* 19, no.1(Spring):83—110.

Rosenstone, Steven R. 1983. *Forecasting Presidential Elections*. New Haven: Yale University Press.

Roth, Alvin E. 1988. "Laboratory Experimentation in Economics: A Methodological Overview." *The Economics Journal* 98(December):974—1031.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688—701.

——. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6:34—58.

Russett, Bruce. 1978. "The Marginal Utility of Income Transfers to the Third World." *International Organization* 32, no.4:913—928.

Sanday, Peggy Reves. 1983. "The Ethnographic Paradigm(s)." In John Van

Maanen, ed. *Qualitative Methodology*. Sage: Beverly Hills.

Schumpeter, Joesph A. [1936] 1991. "Can Capitalism Survive?" In Richard Swedberg, ed. *The Economics of Sociology and Capitalism*, Princeton: Princeton University Press.

Shepsle, Kenneth A. 1986. "Institutional Equilibrium and Equilibrium Institutions." In Herbert F. Weisberg, ed. *Political Science: The Science of Politics*. New York: Agathon Press.

Shively, W. Phillips. 1990. *The Craft of Political Research*. 3d ed. Englewood Cliffs, N.J.: Prentice-Hall.

Simon, Herbert A. 1985. "Human Nature in Politics: The Dialogue of Psychology with Political Science." *American Political Science Review*. 79, no. 2 (June): 293—305.

Skocpol, Theda. 1979. *States and Social Revolutions*. Cambridge University Press.

——. 1984. "Emerging Agendas and Recurrent Strategies in Historical Sociology." In Theda Skocpol, ed. *Vision and Method in Historical Sociology*. New York: Cambridge University Press.

Snyder, Glenn H., and Paul Diesing. 1977. *Conflict among Nations: Bargaining, Decision Making and System Structure in International Crises*. Princeton: Princeton University Press.

Snyder, Jack. 1991. *Myths of Empire: Domestic Politics and International Ambition*. Ithaca: Cornell University Press.

Sober, Elliot. 1988. *Reconstructing the Past: Parsimony, Evolution and Inference*. Cambridge: MIT Press.

Suppes, Patrick C. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Tawney, R. H. 1935. *Religion and the Rise of Capitalism*. New York: Harcourt, Brace & Co.

Tilly, Charles, ed. 1975. *The Formation of National States in Western Europe*. Princeton: Princeton University Press.

Verba, Sidney. 1967. "Some Dilemmas of Political Research." *World Politics* 20 (October): 111—128.

Verba, Sidney, Kay L. Schlozman, Henry Brady, and Norman Nie. 1993. "Race, Ethnicity, and Political Resources: Participation in the United States." *British Journal of Political Science* 23: 453—497.

Waltz, Kenneth N. 1979. *Theory of International Politics*. Reading, Mass: Addison-Wesley.

Webb, Eugene J., D. T. Campbell, R. D. Schwartz, and L. Sechrest. 1966. *Unobtrusive Measures*. Chicago: Rand McNally.

Webb, Eugene J., and Karl E. Weick. 1983. "Unobtrusive Measures in Organizational Theory: A Reminder." In John Van Maanen, ed. *Qualitative Methodology*. Sage: Beverly Hills.

Weber, Max. [1905] 1949. "Critical Studies in the Logic of the Cultural Sciences." In Max Weber, ed. *The Methodology of the Social Sciences*. Translated and edited by Edward A. Shils and Henry A. Fluch. New York: Free Press.

Weiner, Myron. 1991. *The Child and the State in India*. Princeton: Princeton University Press.

Wendt, Alexander. 1992. "Anarchy is What States Make of It: The Social Construction of Power Politics." *International Organization* 64, no. 2 (Spring): 391—426.

Wolfinger, Raymond, and Steven Rosenstone. 1980. *Who Votes*. New Haven: Yale University Press.

Woods, John, and Douglas Walton. 1982. *Argument: The Logic of the Fallacies*. New York: McGraw-Hill Ryerson Ltd.

Zelditch, Morris Jr. 1971. "Intelligible Comparisons." In Ivan Vallier, ed. *Comparative Methods in Sociology*. Berkeley and Los Angeles: University of California Press.

Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

——. 1984. *Basic Issues in Econometrics*. Chicago: University of Chicago Press.